

Throughput-Centric Wave-Pipelined Interconnect Circuits for Gigascale Integration

A Thesis
Presented to
The Academic Faculty

by

Vinita Deodhar

In Partial Fulfillment
of the Requirement for the Degree of
Doctor of Philosophy in Electrical and Computer Engineering



School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2005

Copyright © 2005 by Vinita Deodhar

Throughput-Centric Wave-Pipelined Interconnect Circuits for Gigascale Integration

Approved by:

Dr. Jeffrey Davis, Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. James Meindl
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. D. Scott Wills
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Madhavan Swaminathan
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Paul Kohl
School of Chemical and Biomolecular
Engineering
Georgia Institute of Technology

Date approved: October 27, 2005

Dedicated to
Sri Sri Ravi Shankar, or “Guruji” as we call him ...

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude toward my PhD advisor, Dr. Jeff Davis. It is his intelligent guidance, undying support, and infinite patience that have got me here. I deeply respect him for his exemplary role as a friend, philosopher, and guide, which made working toward my PhD one of the most memorable experiences in my life.

I am grateful to Dr. James Meindl, Dr. Scott Wills, Dr. Madhavan Swaminathan, and Dr. Paul Kohl for their invaluable guidance in my research. I would like to express sincere gratitude to late Dr. John Uyemura for his impeccable teaching skills. I also gratefully acknowledge the support of the National Science Foundation for this work.

I would like to thank my fellow researchers, Pranav, Harshit, Heather, and Gerald, who painted my Georgia Tech days with colors of fun and friendship. The technical discussions about anything in Science, fun-filled birthday celebrations, and numerous eat-outs with them are the memories I will cherish for years to come.

I cannot find enough words to express my gratitude to my family. Their love and support across the seven seas is what keeps me going. Finally, I would like to take this opportunity to thank my fiancé and fellow researcher, Ajay, whose invaluable support on the technical and emotional fronts made this journey one of the most special experiences in our relationship. It is the unconditional love of my parents and my fiancé that gives a meaning to simple things in my life.

TABLE OF CONTENTS

Dedication	i
Acknowledgements	ii
List of tables	ix
List of figures	xiii
Summary	xvii
 1. Introduction and background	 1
1.1 Solutions to the interconnect problem: performance, power, and area.....	3
1.1.1 Repeater insertion for performance enhancement of on-chip interconnects.....	3
1.1.2 Power reduction techniques for on-chip interconnects.....	6
1.1.3 Wire-size optimization of interconnect networks.....	8
1.1 Proposed research.....	10
1.2.1 Research objective.....	10
1.2.2 Summary of chapters.....	11
 2. Performance models for wave-pipelined interconnects	 14
2.1 Introduction.....	14
2.2 Concept of wave-pipelining.....	15
2.3 Derivation and validation of analytical throughput model for wave-pipelined RC interconnects.....	18
2.3.1 Derivation of analytical throughput model.....	18
2.3.2 Validation of analytical throughput model.....	25
2.3.3 Multisegment coexistence of data.....	29
2.3.4 Saturation throughput.....	31
2.4 Impact of scaling on maximum wire bit rate.....	35
2.4.1 Length scaling.....	35
2.4.2 Constant field scaling.....	38

2.4.3	Interconnect scaling.....	39
2.4.4	Transistor width scaling.....	42
2.5	Impact of changes in design assumptions on throughput.....	45
2.5.1	The last repeated wire segment achieves a normalized voltage swing other than 0.9.....	46
2.5.2	The inverters in a repeater have a normalized switching threshold other than 0.5.....	48
2.5.3	A repeater consists of a single inverter instead of two.....	51
2.6	Summary.....	54
3.	Signal integrity analysis of wave-pipelined interconnects	55
3.1	Introduction.....	55
3.2	Performance models for transmission line.....	56
3.2.1	RLC bandwidth model.....	57
3.2.2	RLC time delay model.....	59
3.2.3	RLC rise time model.....	61
3.3	Boundary between RC and RLC models.....	63
3.4	Comparison of bit rates obtained on transmission line and wave-pipelined interconnect.....	67
3.5	Impact of wave-pipelining on RLC crosstalk.....	70
3.5.1	Simulation of RLC interconnect circuits using HSPICE and RAPHAEL.....	70
3.5.2	Impact of wave-pipelining on dynamic delay effects, overshoot voltage, and crosstalk voltage in RLC interconnect systems.....	72
3.6	Impact of wave-pipelining on power-supply noise.....	77
3.7	Techniques to minimize performance variations on wave-pipelined interconnect circuits.....	80
3.7.1	Shielding ground lines insertion.....	80
3.7.2	Misaligned repeater insertion.....	81
3.7.3	Decoupling capacitor insertion.....	83
3.8	Summary.....	84

4. Voltage scaling repeater insertion (VSRI) circuit analysis	85
4.1 Introduction.....	85
4.2 Importance of voltage scaling.....	86
4.3 Simultaneous application of voltage scaling and repeater insertion (VSRI).....	88
4.4 Impact of VSRI on signal integrity.....	91
4.4.1 Impact of VSRI on overshoot voltage.....	92
4.4.2 Impact of VSRI on crosstalk voltage.....	92
4.5 Comparison of VSRI and LVDS.....	94
4.5.1 Circuit configuration for LVDS and VSRI.....	94
4.5.2 LVDS and VSRI comparison for performance, power, and area.....	95
4.5.3 Impact of power supply noise on LVDS and VSRI	96
4.5.4 VSRI overdesign for stable performance.....	101
4.6 Summary.....	105
5. Voltage scaling, repeater insertion, and wire sizing optimization	106
5.1 Introduction.....	106
5.2 Voltage scaling, repeater insertion, and wire sizing.....	107
5.3 Design metrics for different types of applications.....	108
5.3.1 Low-power, high-performance applications.....	108
5.3.2 Moderate-power, moderate-performance, area-constrained applications.....	118
5.3.3 Ultra-high-performance applications.....	123
5.4 Comparison of optimal design points.....	126
5.5 Comparison of LVDS and TPEA optimization.....	129
5.6 Impact of interconnect geometries on different design optimizations.....	133
5.7 Impact of simultaneous application of voltage scaling, repeater insertion, and wire sizing on via blockage.....	135
5.8 Latency-sensitive wave-pipelining.....	137
5.8.1 Optimal latency-centric wave-pipelining.....	140
5.8.2 Wave-pipelining with suboptimal repeater insertion.....	142
5.9 Summary.....	144

6. Synchronization of data on wave-pipelined interconnects	146
6.1 Introduction.....	146
6.2 Comparison of wave-pipelining and latch insertion.....	147
6.2.1 Wave-pipelined interconnect with simple latch receiver.....	147
6.2.2 Latch-inserted interconnect.....	148
6.2.3 Performance of latch-inserted interconnect and wave-pipelined interconnect with receiver.....	150
6.2.4 Comparison of latch-inserted interconnect with wave-pipelined interconnect under a constant throughput constraint.....	151
6.3 Existing solutions for synchronizing data on wave-pipelined interconnects.....	153
6.3.1 Sending clock along with data and latching data using flip-flops.....	153
6.3.2 Locally generating clock and latching data using PLL.....	156
6.4 Simplified receiver for fully synchronous systems.....	157
6.4.1 Interfacing DDR cores with wave-pipelined DDR interconnects.....	158
6.4.2 Interfacing SDR cores with wave-pipelined DDR interconnects.....	161
6.5 Globally asynchronous locally synchronous (GALS) systems.....	163
6.5.1 Skew-insensitive retimer circuit.....	165
6.5.2 Generation of control signals for retimer circuits.....	168
6.5.3 HSPICE simulation of wave-pipelined interconnect with retimer circuit.....	171
6.5.4 Area and power for wave-pipelined interconnect circuit.....	172
6.5.5 Retimer circuit for SDR driver and receiver cores.....	173
6.6 Summary.....	176
7. Impact of technology scaling on wave-pipelining	177
7.1 Introduction.....	177
7.2 Evaluation of transistor parameters.....	178
7.2.1 Evaluation of transistor resistance.....	178
7.2.2 Evaluation of transistor capacitance.....	180
7.3 Evaluation of interconnect parameters.....	181
7.3.1 Evaluation of interconnect resistance.....	182
7.3.2 Evaluation of interconnect capacitance.....	183

7.3.3	Evaluation of interconnect inductance.....	184
7.4	Comparison of throughput using analytical expression and HSPICE for a 45 nm node.....	184
7.5	Impact of technology scaling on communication throughput.....	187
7.5.1	ITRS projections for on-chip local clock frequency.....	187
7.5.2	Impact of technology scaling on throughput for two different design scenarios.....	188
7.5.3	Future of latency-centric repeater insertion.....	190
7.5.4	Impact of technology scaling on power and performance of global interconnects.....	191
7.5.5	Impact of technology scaling on supply voltage.....	193
7.6	Performance, power, and area analysis of 32 nm node.....	194
7.6.1	Design optimization for 32 nm node.....	194
7.6.2	Impact of via area on design optimization.....	197
7.6.3	Power breakdown for a 32 nm global interconnect circuit.....	198
7.6.4	Multilayer interconnect throughput for 32 nm node.....	199
7.7	Impact of material alternatives on wave-pipelining.....	201
7.7.1	High resistivity resulting from scattering.....	201
7.7.2	Difficulties in achieving low permittivity for interlayer dielectric	203
7.7.3	Insignificant mobility enhancement and large oxide thickness.....	204
7.8	Summary.....	205
8.	Conclusions and future work	207
8.1	Future work.....	207
8.1.1	Future work: improvement of physical and analytical models.....	207
8.1.2	Future work: analysis of manufacturing and process variations.....	208
8.1.3	Future work: analysis of wave-pipelining for different wiring net models.....	209
8.1.4	Future work: extension of system-level analysis.....	210
8.2	Salient conclusions of dissertation.....	211
	Appendix A. Estimation of transistor resistance.....	216

Appendix B. Waveforms at different nodes for a wave-pipelined interconnect.....	218
Appendix C. Values of design metrics for different design choices.....	220
References.....	237
List of publications.....	244
Vita.....	245

LIST OF TABLES

1.1	Intel interconnect dimensions for a multilayer architecture in 180 nm technology.....	9
2.1	Summary of design equations for the analytical throughput model.....	24
2.2	Values of v_1 for up to 50 repeaters.....	25
2.3	Physical parameters used for 180 nm HSPICE simulations.	26
2.4	Comparison of bit rates using different models.	28
2.5	Values of N_s for different repeater densities.....	30
2.6	Values of l_{eff} for different repeater densities.	31
2.7	Saturation throughput using (2.30) for various supply voltages.	33
2.8	Summary of impact of scaling on maximum throughput.....	45
2.9	Values of v for different normalized voltage swings at the output of the last segment (over 20 iterations).....	48
2.10	Values of inverter threshold voltage for two different supply voltages.....	49
2.11	Values of v for different inverter threshold voltages (over 20 iterations).....	51
2.12	Comparison of throughput for non-inverting and inverting repeaters.....	54
3.1	Dimensions and parasitics for transmission line.	56
3.2	The comparison of delay calculated using models in [51] and HSPICE simulations.....	61
3.3	Values of Z_0' for different repeater densities.....	65
3.4	Comparison between performance of RC and RLC wave-pipelined interconnect circuits using HSPICE simulations.....	66
3.5	Impact of input resistance on transmission line performance.....	67
3.6	Transmission line and wave-pipelining design choices for a constant throughput performance.....	69
3.7	Parameters used in HSPICE simulations for crosstalk analysis.....	74
3.8	HSPICE Results for throughput, latency, and active line overshoot voltage.....	75
3.9	Nominal values of throughput and latency for low-loss transmission line and wave-pipelined interconnect circuit.....	78

3.10	Absolute errors for low-loss transmission line and wave-pipelined interconnect.....	79
3.11	HSPICE results for a 180 nm global interconnect with four repeaters.	82
4.1	Active line overshoot voltage normalized to supply voltage for different supply voltages and repeater densities.....	92
4.2	Quiet line crosstalk voltage normalized to supply voltage for different supply voltages and repeater densities.....	93
4.3	LVDS and VSRI circuit configurations for a 0.5 cm long 180 nm interconnect...	94
4.4	Comparison between LVDS and VSRI for a 2 Gbps throughput.....	95
4.5	Average and maximum values of absolute error for LVDS and VSRI	98
4.6	Absolute errors for a VSRI design point of 5 repeaters per 0.5 cm.....	104
4.7	Comparison between new VSRI design and existing LVDS design.....	104
5.1	Values of θ and corresponding $V_{dd,opt}$	112
5.2	Summary of design parameters for the optimal TPBE design point.....	116
5.3	Optimal TPBE design point.....	117
5.4	Two design choices to achieve an aggregate throughput of 6 Gbps.....	120
5.5	Optimal TPEA design point.....	122
5.6	Optimal TPA design point.....	126
5.7	Optimal design points for different applications.....	127
5.8	Interconnect geometries and other parameters corresponding to different optimal design points.	127
5.9	Optimal latency-centric design point.	128
5.10	Comparison of optimal design points for a constant throughput of 1.25 Gbps...	129
5.11	Comparison of LVDS and TPEA optimization for a 0.5 cm link.....	132
5.12	Design optimizations for two different geometries.....	134
5.13	Comparison of latency-centric and throughput-centric design approaches in terms of via blockage, power, area.....	136
5.14	Latency-centric repeater insertion and wave-pipelining.....	141
6.1	Performance comparison of wave-pipelined interconnect with latch-inserted interconnect in 180 nm technology generation.....	150

6.2	Comparison of latch insertion and wave-pipelining for a constant throughput of 3 Gbps.....	152
6.3	Values of control signals when CLK_R is in phase with CLK_S.....	167
6.4	Design for skew tolerance of 0° to 360° between CLK_S and CLK_R.....	168
6.5	Design details for wave-pipelined interconnect circuit.....	173
7.1	Technology parameters for different technology generations [2].....	178
7.2	Transistor resistance and other parameters for a transistor scaling factor of 56, for different technology generations.....	179
7.3	Transistor capacitance for a transistor scaling factor of 56, for different technology generations.....	180
7.4	Global interconnect dimensions for different technology generations.....	182
7.5	Interconnect resistance per unit cm for different technology generations.....	183
7.6	Interconnect capacitance per unit cm for different technology generations.....	183
7.7	Interconnect inductance per unit cm and coupling factors for different technology generations.....	184
7.8	Physical parameters used for 45 nm HSPICE simulations.....	185
7.9	Comparison of bit rates using reciprocal latency and throughput for 45 nm global interconnects.....	186
7.10	On-chip local clock frequency for different technology generations [2].....	187
7.11	Optimal values of parameters for latency-centric repeater insertion on a 1 cm long interconnect.....	190
7.12	Power and performance of a 1 cm long global interconnect to achieve the projected throughput for various technology generations.....	192
7.13	A 32 nm interconnect system with 12 metal levels.....	194
7.14	Design parameters for a 1 cm long 32 nm global interconnect routed on metal-12.....	195
7.15	Performance, power, and area for a 1 cm long 32 nm global interconnect.....	196
7.16	Impact of via area on design metrics.....	197
7.17	Values of resistance and capacitance for a 1 mm long 32 nm metal-6 interconnect.....	201
A.1	Values of R_t from I-V curve and approximations.....	217

C.1	Values of TPBE in Gbps/pJ in design space.....	221
C.2	Values of TPEA in Tbps/pJ-cm ² in design space.....	225
C.3	Values of TPA in Tbps/cm ² in design space.....	229
C.4	Values of latency in ns in design space.....	233

LIST OF FIGURES

1.1	Impact of technology scaling on gate delay and interconnect delay [4].....	2
1.2	An RC interconnect with repeaters.....	4
1.3	Current evolution of interconnect architecture for high-performance CMOS systems.....	9
2.1	Difference between the values of throughput and reciprocal latency for a supply voltage of 2 V.....	17
2.2	RC model for wave-pipelined interconnect circuit.....	19
2.3	Interconnect cross-sectional dimensions.....	20
2.4	Waveforms at the input and output of k^{th} repeater.....	22
2.5	Values of v at the output of every repeated segment for an interconnect with 50 repeaters.....	23
2.6	Curve-fitting for values of v_l for up to 50 repeaters.....	25
2.7	Comparison between results of closed-form analytical expression and HSPICE simulations.....	27
2.8	Comparison of bit rates using (2.16) and (2.17) with HSPICE simulations.....	29
2.9	Variation of throughput with the repeater density for different interconnect lengths.....	37
2.10	Variation of throughput with number of repeaters for different square cross-sections for a 2 V supply, using analytical throughput model.....	40
2.11	Variation of throughput with interconnect width for constant pitch for different interconnect aspect ratios, using (2.16).....	42
2.12	An RC interconnect with inverting repeaters.....	52
3.1	Transmission line structure.....	56
3.2	Transmission line and its equivalent electrical circuit model in [50].....	57
3.3	Comparison of analytical expression in (3.13) and HSPICE simulation results for transmission line delay with finite input rise time.....	62
3.4	A 5 interconnect, 2 ground plane system with self and mutual capacitances.....	71
3.5	A 5 interconnect system with self and mutual components of inductance.....	72

3.6	Equivalent capacitances and inductances for different switching patterns for a 5 interconnect, 2 ground plane system.....	73
3.7	Normalized crosstalk voltage on a quiet line.....	77
3.8	Histograms for throughput and latency of transmission line.....	78
3.9	Histograms for throughput and latency of wave-pipelined interconnect.....	79
3.10	Interconnects with misaligned repeaters [5].....	82
4.1	Circuit implementation of LVDS [18]	87
4.2	Effect of VSRI on throughput.....	90
4.3	Effect of VSRI on communication latency.....	91
4.4	Example of supply noise with maximum frequency component of 4 GHz.....	97
4.5	Histograms for values of throughput for LVDS and VSRI, over 1000 simulations.....	99
4.6	Histograms for values of latency for LVDS and VSRI, over 1000 simulations.....	100
4.7	Histograms for throughput and latency for a VSRI design point of 5 repeaters per 0.5 cm interconnect length.....	103
5.1	Variation of TPBE with supply voltage.....	113
5.2	Variation of TPBE with number of repeaters on an interconnect with a 250 nm square cross-section, for different supply voltages.....	114
5.3	Interconnect dimensions.....	117
5.4	Variation of TPEA with number of repeaters on an interconnect with 250 nm square cross-section, for different supply voltages.....	122
5.5	Effect of repeater density and wire dimensions on throughput performance.....	124
5.6	Variation of TPA with number of repeaters on an interconnect with 250 nm square cross-section, for a 2 V supply.....	125
5.7	Two different design choices for 1.3 Gbps SPARC memory bus.....	130
5.8	Different design geometries for an interconnect system.....	134
5.9	Five-stage MIPS pipeline.....	139
5.10	Bypass bus in pipelined partial datapath.....	140
5.11	Flat minima for latency of a 180 nm interconnect.....	143
5.12	Comparison of throughput by latency-centric approach and wave-pipelining ...	144

6.1	Wave-pipelined interconnect with receiver.....	148
6.2	Latch-inserted interconnect.....	149
6.3	Schematic representation of latch-inserted interconnect and wave-pipelined interconnect with receiver.....	151
6.4	Flip-flop-based receiver in [5].....	154
6.5	PLL-based receiver in [16]	156
6.6	Receiver for wave-pipelined interconnect connecting DDR cores.....	158
6.7	Circuit diagram for 2:1 multiplexer in 180 nm technology.....	160
6.8	Timing waveform for DDR cores interfaced with DDR interconnect.....	161
6.9	Interfacing SDR cores with wave-pipelined DDR interconnect.....	162
6.10	Timing waveform for SDR cores interfaced with DDR interconnect.....	163
6.11	Circuit schematic for DDR interconnects with DDR retimer for GALS systems.....	165
6.12	Retimer circuit for wave-pipelined interconnects in GALS scenario.....	166
6.13	Waveforms for S_1S_0 and R_1R_0	168
6.14	Simple AND gates used for generating STROBE signals.....	169
6.15	Toggle latches for generation of control signals.....	170
6.16	Waveforms for various data and control signals in retimer circuit.....	172
6.17	Clock and control signals for SDR driver-receiver cores.....	174
6.18	A generalized 2-stage toggle latch for generating control signals for SDR driver- receiver cores in GALS scenario.....	175
6.19	Waveforms for retimer circuit for SDR driver-receiver cores.....	175
7.1	Values of global wire pitch for present technology generations and their extrapolation for future technology generations.....	182
7.2	Comparison of throughput using analytical model and HSPICE simulations for a 1 cm long 45 nm interconnect.....	186
7.3	Throughput using optimal-sized repeaters for different technology generations.....	189
7.4	Comparison of reciprocal latency and throughput requirement for various technology generations.....	191

7.5	Comparison of ratios of supply voltage to threshold voltage by ITRS [2] and low-power design in this research, for different technology generations.....	193
7.6	Power breakdown for a 32 nm global interconnect circuit.....	198
7.7	Throughput of a 1 cm interconnect in 32 nm technology generation.....	199
7.8	Throughput of a 1 mm interconnect in 32 nm technology generation.....	200
7.9	Repeater density to achieve a 23 Gbps throughput on different metal levels in a 32 nm node.....	203
7.10	Repeater density to achieve a 23 Gbps throughput on metal-10 in a 32 nm node for different values of relative dielectric permittivity.	204
8.1	A multiple-source, multiple-sink wiring net model.....	210
A.1	I-V curve for a 180 nm NMOS.	217
B.1	HSPICE results for voltages at the input, first repeated segment, and output of a wave-pipelined interconnect having a 250 nm x 250 nm cross-section.....	219
C.1	Interconnect dimensions.	220

SUMMARY

The central thesis of this research is that VLSI interconnect design strategies should shift from using global wires that can support only a single binary transition during the latency of the line to global wires that can sustain multiple bits traveling simultaneously along the length of the line. It is shown in this thesis that such throughput-centric multibit transmission can be achieved by wave-pipelining the interconnects using repeaters. A holistic analysis of wave-pipelined interconnect circuits, along with the full-custom optimization of these circuits, is performed in this research. With the help of models and methodologies developed in this thesis, the design rules for repeater insertion are crafted to simultaneously optimize performance, power, and area of VLSI global interconnect networks through a simultaneous application of voltage scaling and wire sizing. A qualitative analysis of latency, throughput, signal integrity, power dissipation, and area is performed that compares the results of design optimizations in this work to those of conventional global interconnect circuits. The objective of this thesis is to study the circuit- and system-level opportunities of voltage scaling, wire sizing, and repeater insertion in wave-pipelined global interconnect networks that are implemented in deep submicron technologies.

CHAPTER 1

INTRODUCTION AND BACKGROUND

The rapid advancement of computing technology in the twentieth century has been phenomenal. This phenomenon began in 1948 with Bardeen, Brattain, and Shockley's invention of solid-state transistor. As a result of this invention coupled with Kilby and Noyce's invention of integrated circuit technology in 1958, computing technology has advanced exponentially over the past few decades. Through the fullest exploitation of the latest technology transistors and interconnects, Intel's next-generation billion-transistor Itanium microprocessor has set out to achieve an industry-leading performance of 1 GHz [1]. Interestingly, the processing power has moved from performing simple calculations to managing complex billion-transistor systems in a relatively short period of time. With the invention of new manufacturing technologies and continuous transistor scaling, large IBM mainframes that once occupied several rooms can now fit into a small area of a few square centimeters.

The unrelenting scaling of CMOS technology results in rapid changes in the power and performance trends of very large scale integrated (VLSI) systems and requires continuous changes to the circuit design techniques and system architectures. In an effort to outline the limits and opportunities for the future CMOS technology nodes, a consortium of semiconductor industry associations around the world created the

International Technology Roadmap for Semiconductors (ITRS). According to the latest (2004) ITRS update, the number of transistors per chip is expected to reach well above one billion by the end of this decade [2]. The number of metal layers is expected to increase in proportion, and a multifold increase in the internal clock frequency is also projected. Therefore, supporting high-performance applications on multibillion transistor chips with an acceptable number of metal levels and power dissipation is one of the biggest challenges before VLSI designers.

The total delay (i.e., latency) of a circuit is traditionally considered to be the measure of its performance. The total delay comprises two components, the transistor delay and the interconnect delay. As seen in Figure 1.1, with the scaling of CMOS technology, the global interconnect delay that was once significantly smaller than the transistor delay has now become a few hundred times larger than the transistor delay [3]. As a result, on-chip interconnects are limiting the maximum performance that can be achieved on processor systems.

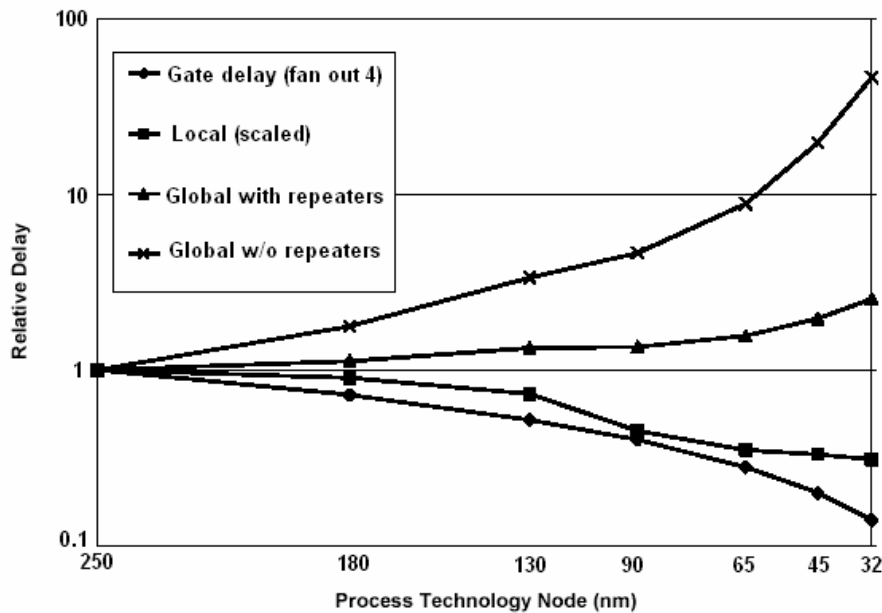


Figure 1.1: Impact of technology scaling on gate delay and interconnect delay [4].

With the advent of deep submicron (DSM) technology, a significant amount of functionality is being integrated onto a single chip such that long global interconnects have become a common on-chip feature [5]. Identifying that the improvement in speed to support the continuously increasing speed of microprocessors must come from these interconnects, a lot of time and effort are being directed toward the analysis of on-chip global interconnects. Moreover, because of the importance of power dissipation and chip size in both high-performance and portable applications, it is necessary to balance both these parameters and still achieve a high interconnect performance to support the high-speed applications.

1.1 Solutions to the interconnect problem: performance, power, and area

1.1.1 Repeater insertion for performance enhancement of on-chip interconnects

Having identified interconnects as the primary performance bottlenecks, several techniques have been proposed to enhance their performance. In 1985, Bakoglu and Meindl presented the repeater insertion technique to enhance the interconnect performance. In this technique, a long interconnect is divided into several smaller segments, and each segment is driven by a repeater. The repeater circuits can be as simple as a single inverter or a pair of inverters. A resistance-capacitance (RC) model of an interconnect with repeaters is shown in Figure 1.2.

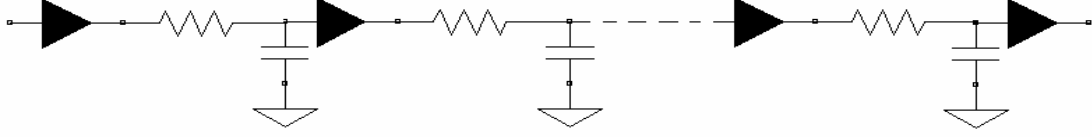


Figure 1.2: An RC interconnect with repeaters.

Bakoglu and Meindl derived the time delay expression for a repeater-inserted interconnect and calculated the optimal number and optimal size of repeaters that minimize the total interconnect delay [6], [7]. The expression for the optimal number of repeaters is given as

$$n_{opt} = \sqrt{\frac{RC}{2.3R_0C_0}}, \quad (1.1)$$

and that for the optimal size of repeaters is given as

$$h_{opt} = \sqrt{\frac{R_0C}{C_0R}}, \quad (1.2)$$

where R_0 and C_0 are the resistance and capacitance of a minimum-sized transistor and R and C are the resistance and capacitance of the interconnect, respectively. At this optimal design point, the 50% delay is given by

$$T = 2.46\sqrt{R_0C_0RC}. \quad (1.3)$$

The delay of a single-driver RC interconnect is dominated by the product of R and C , which vary directly with the interconnect length. Therefore, the total delay varies with the square of the interconnect length when the interconnect is driven by a single driver. However, it is seen in (1.3) that for an interconnect with the optimal number and optimal

size of equal-spaced repeaters, the interconnect delay, which is minimized at this optimal design point, varies directly with the interconnect length. Thus, though it may seem counterintuitive, it is shown in [6] that adding repeaters can result in a lower interconnect delay than that of a single-driver interconnect.

Several variations of the repeater insertion technique, analyzed from various design approaches, can be found in the literature. For instance, the suboptimal repeater insertion technique that tolerates a slightly larger delay is suggested in [8]-[10] to reduce silicon area. An optimal repeater insertion technique to reduce the total power dissipation is explored in [11], [12]. Adler et al. have analyzed uniform repeater insertion and tapered-buffer repeater insertion in [13] to reduce delay and power on the interconnects and have found that uniform repeater insertion outperforms tapered-buffer insertion.

The misaligned repeater insertion technique, along with interleaved lines, is discussed in [5] to improve the interconnect performance and signal integrity in a network-on-chip (NoC). This technique is further optimized in [14] to minimize the propagation delay of the interconnect circuit. The insertion of first-in-first-out (FIFO) buffers on the interconnects is suggested in [15] to handle different data rates between IP cores in a system-on-chip (SoC).

However, most repeater insertion techniques focus on minimizing the interconnect latency because latency is considered to be the primary measure of interconnect performance. For traditional VLSI designs, optimal repeater insertion suggested in [6] maximizes the throughput performance, which is the inverse of the latency. However, for the repeater-inserted interconnects, the throughput need not be restricted to the reciprocal latency. There is an opportunity to substantially increase the

communication throughput of repeater-inserted interconnects beyond the reciprocal latency through high-speed serialization of the data, and this is achieved by wave-pipelining. The use of wave-pipelining to enhance the throughput performance of repeater-inserted interconnects is discussed in [5], [16]. The authors of [5] and [16] also present timing analyses of wave-pipelined interconnects using different receiver circuits to capture the data at the output of interconnects.

1.1.2 Power reduction techniques for on-chip interconnects

As a result of a large number of transistors per chip and a high operational frequency, ITRS projects high values of power dissipation for present and future processor generations [2]. An increased transistor count along with an increase in design complexity also results in a proportional increase in the number of interconnects on the chip, and the interconnect power becomes a significant portion of the total power dissipation on the chip [17]. Because of the increased parasitic capacitance, repeater insertion further increases the total power dissipation on interconnect circuits, which necessitates the use of low-power techniques for interconnects.

The dynamic power is the primary contribution to the total power for technology generations up to 100 nm. Because the dynamic power varies with the square of the voltage swing that a circuit undergoes, several low-voltage-swing techniques are proposed by researchers to reduce dynamic power. Low-voltage differential signaling (LVDS) is one signaling standard used in the industry that was developed to obtain high speed on low-power interconnects. In this technique, a reduced voltage swing of a few hundreds of milli-volts is used on the differential interconnect [18] to reduce the

interconnect power. Some other low-voltage-swing techniques such as pulse-controlled drivers [18] and symmetric driver and level converters for low-power ULSIs [19] are also found in the literature.

The dynamic power on interconnects also varies directly with the switching activity. Reducing switching activity can proportionally reduce the interconnect power. Therefore, some researchers have proposed techniques such as bus invert coding [20] and pulsed-wave interconnect [21] that attempt to reduce the interconnect power by using encoding techniques to minimize the switching activity on the interconnect.

Even though the contributions of the leakage power (resulting from subthreshold and gate leakage) and the short-circuit power are negligible for technology generations up to 100 nm, they are expected to be significant portions of the total power for future technology generations [12]. Though dynamic power is the only possible type of power dissipation on *interconnects*, with the insertion of repeaters for performance enhancement, the consideration of leakage power and short-circuit power becomes important for the *interconnect circuits*.

Reducing the voltage swing on interconnects reduces the interconnect dynamic power and reducing the switching activity reduces dynamic and short-circuit power, but because the transistors in driver-receiver circuits operate from the full-swing supply voltage, they still dissipate large amounts of power. Moreover, the total power dissipation in the additional circuitry that is used to obtain lower swings on interconnects or perform bus encoding could counteract the power reduction on interconnects, thereby making such techniques viable only for large interconnect lengths.

In the systems using a single power supply voltage, the dynamic power varies with the square of the supply voltage and the leakage and short-circuit power vary directly with the supply voltage. Therefore, the scaling of the supply voltage is essential to reduce the *total* power dissipation. Supply voltage scaling and some other techniques to reduce total power (e.g., frequency scaling, using larger threshold voltages) are discussed in [22]. To discover techniques to reduce the leakage and the short-circuit power for future technology generations, accurate power modeling is needed. The models for the leakage power are presented in [12], [22]-[24] and those for the short-circuit power are presented in [12], [25], [26].

1.1.3 Wire-size optimization of interconnect networks

According to the ITRS projections, the number of metal levels is expected to reach 12 by the end of the current decade. Figure 1.3 shows a current evolution of the interconnect architecture for high-performance CMOS systems [27]. Because the manufacturing cost is directly proportional to the number of metal levels [28], focusing on the techniques that reduce wire area is as important as developing low-power techniques. Optimizing the number and dimensions of interconnects is critical for an efficient utilization of available wiring tracks in a multilevel interconnect stack. Table 1.1 shows Intel dimensions for a multilayer interconnect architecture in the 180 nm technology [29]. It is seen in Table 1.1 that because metal-5 and metal-6 typically route global interconnects, they are reverse-scaled to have a larger wire pitch.

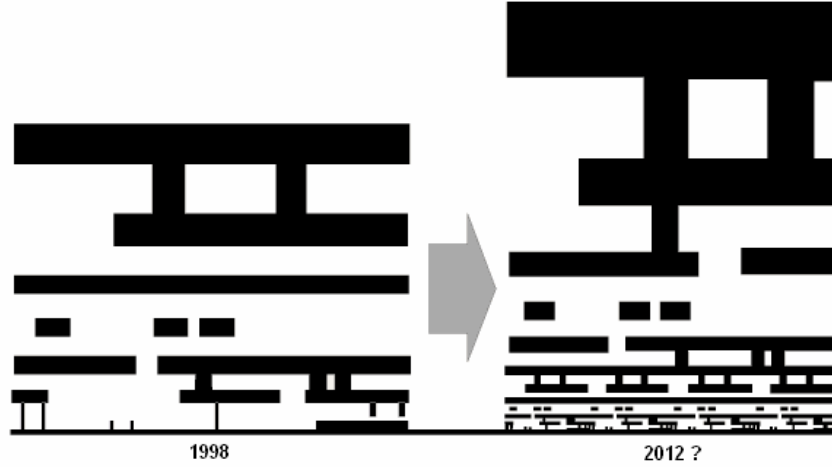


Figure 1.3: Current evolution of interconnect architecture for high-performance CMOS systems.

Table 1.1: Intel interconnect dimensions for a multilayer architecture in 180 nm technology [29].

Metal level	Wire pitch (nm)	Metal height (nm)	Aspect ratio
1	500	480	1.9
2	640	700	2.2
3	640	700	2.2
4	1080	1080	2.0
5	1600	1600	2.0
6	1720	1720	2.0

Various schemes have been proposed to optimize interconnect dimensions and wire layer assignments [30]-[33]. Venkatesan et al. have presented a holistic multilevel interconnect architecture design methodology that has been integrated into a multilevel interconnect network design simulator ‘MINDS’, which optimizes the interconnect cross-sectional dimensions of each metal layer to reduce logic macrocell area, cycle time, power consumption, and number of metal layers [34]. The MINDS simulator is further developed in [28] by the inclusion of HSPICE to give more accurate results for the number of metal levels.

Naeemi et al. have studied the impact of wire width on latency and have identified the optimal wire width that is independent of interconnect length to offer the best trade-off between latency and data flux density [35]. The data flux density in [35] is given by

$$\Phi_D(w) = \frac{\text{Bandwidth}}{\text{pitch}} = \frac{\left(\frac{1}{\text{latency}}\right)}{2 \cdot \text{width}}, \quad (1.4)$$

whose integration over the chip edge gives the total bisectional bandwidth, i.e., the number of bits that cross the central bisectional line of length equal to the chip edge.

Because dynamic delay and crosstalk effects resulting from inductive and capacitive coupling with adjacent interconnects are a function of interconnect dimensions, optimizing interconnect dimensions is also important to maintain good signal integrity. Compact models for interconnect delay and crosstalk are rigorously derived in [36]-[38]. Using the compact crosstalk models in [37], different geometries of RLC interconnect systems are analyzed in [35] for different types of switching events. An optimal wire sizing methodology to reduce simultaneous switching noise (SSN) is described in [39].

1.2 Proposed research

1.2.1 Research objective

As seen in the previous section, several variations of the repeater insertion technique to reduce interconnect delay, low-voltage and low-activity techniques to reduce interconnect power for a given performance, and wire-size optimization techniques to reduce interconnect area to meet a certain performance requirement are found in the literature. All of these optimization techniques attempt to minimize the latency because

latency is considered to be the primary measure of performance. However, this research decouples the interconnect *throughput* performance from the latency for the wave-pipelined interconnect circuits and focuses on maximizing the interconnect throughput. Moreover, there have been relatively fewer attempts in the past to optimize both interconnect as well as logic circuit parameters to simultaneously obtain high performance, low power, and low area, which this research attempts to achieve through the simultaneous application of voltage scaling, repeater insertion, and wire sizing.

The objective of the proposed research is to study the circuit- and system-level opportunities of voltage scaling, repeater insertion, and wire sizing in wave-pipelined global interconnect networks that are implemented in DSM technologies. The goal of this thesis is to develop models and methodologies to design global interconnect networks that have low power, low area, and high throughput (i.e., bit rate).

It is shown in this research that repeaters can be effectively used to wave-pipeline the interconnect and enhance its throughput performance significantly beyond the reciprocal of latency. Communication throughput is considered to be a measure of performance in this research, and voltage scaling, repeater insertion, and wire sizing are analyzed from a throughput-centric approach. Various modeling and design methodologies are introduced in this thesis that optimize global interconnect networks to achieve low power, low area, and high throughput performance.

1.2.2 Summary of chapters

The wave-pipelining technique using repeaters is discussed in detail and its importance to enhance the performance of the on-chip global interconnects is explained

in Chapter 2. A closed-form analytical expression is also derived in the same chapter to evaluate the throughput of wave-pipelined RC interconnects. This throughput model is validated using HSPICE and is also used to study the effect of various technology and system parameters on the communication throughput.

Even though the analytical expression calculates the bit rate of the RC interconnects, the impact of inductance on the bit rate and latency of the interconnect is analyzed in Chapter 3. The usefulness of wave-pipelining to achieve high performance and good signal integrity in the presence of severe inductive and capacitive coupling and power supply fluctuations is also discussed in this chapter.

Because scaling of the supply voltage is critical to reduce the total power, the simultaneous application of voltage scaling and repeater insertion (VSRI) is proposed in Chapter 4 for low-power, high-performance interconnects. VSRI is also compared to LVDS in the presence of power supply noise. It is shown that VSRI significantly reduces power, wire area, and latency compared to LVDS, without any loss of throughput performance.

The wire-sizing optimization for the VSRI circuits is discussed in Chapter 5 to minimize the wire area of high-performance interconnects. Using different design metrics, Chapter 5 guides the voltage scaling, repeater insertion, and wire sizing design optimizations for different types of applications. It is important to note that similar to [35], these design optimizations are not based on any assumptions for the wiring distributions and are therefore general, fully scalable, and therefore applicable to interconnects of any length in any technology generation.

Chapters 2-5 present some process-level and primarily circuit-level analysis of the wave-pipelined interconnect circuits. However, the system-level timing analysis of these interconnect circuits is essential for integrating wave-pipelined interconnects into actual processor systems. Therefore, Chapter 6 presents an overview of the existing techniques for correctly capturing the data on wave-pipelined interconnects [5], [16] and presents novel circuit designs to interface wave-pipelining with fully synchronous systems and globally asynchronous locally synchronous (GALS) systems.

The continued scaling of the CMOS technology miniaturizes the transistor channel length by a factor of 0.7 with every new technology node [2]. It is important to study the impact of technology scaling on the performance of the wave-pipelined interconnects to study the viability of wave-pipelining for future interconnect circuits. Based on the performance limits and opportunities outlined by ITRS, Chapter 7 discusses the role of wave-pipelining for performance enhancement of future global interconnect circuits. The analytical throughput model derived in Chapter 2 is effectively used in this chapter to project the future of wave-pipelining.

Finally, Chapter 8 presents the conclusions and salient features of this research and also outlines the possible future work in this area. Some of the future work could involve improving various analytical models, analyzing the impact of manufacturing and process variations on wave-pipelining, and further extending the analysis of wave-pipelining to different wiring net models and multilevel interconnect architectures.

CHAPTER 2

PERFORMANCE MODELS FOR WAVE-PIPELININED INTERCONNECTS

2.1 Introduction

The repeater insertion technique proposed in [6] to minimize the total interconnect latency is discussed in the previous chapter. This chapter explains the wave-pipelining technique using repeaters to enhance the interconnect throughput without a significant degradation of latency. First, the concept of wave-pipelining and its journey from the logic circuits to the interconnect circuits are discussed. A simple closed-form expression is then rigorously derived to evaluate the throughput of the wave-pipelined RC interconnect. The results from this analytical throughput expression are compared to those from HSPICE simulations using RLC interconnect circuits.

Using the analytical throughput model, the effect of various process and system parameters on the interconnect throughput is analyzed. The strength of this simple model is highlighted through its use to study the impact of constant field scaling, interconnect scaling, and transistor scaling on throughput. Finally, the impact of changes in some of the underlying assumptions on the throughput of wave-pipelined interconnects is analyzed, and the application of the model to different design scenarios is discussed.

2.2 Concept of wave-pipelining

Wave-pipelining for logic circuits was proposed in 1969 [40], and this concept has been considerably developed over the last 40 years [41]-[44]. In general, the wave-pipelining technique advocates the application of a new input signal to a combinational logic circuit before the previous input reaches its destination [16]. A combinational circuit consists of several data processing stages. These stages can be considered to be analogous to the stages in a typical data pipeline. Similar to the data pipelines in computer architecture, the main idea behind wave-pipelining is that after a certain processing unit processes a set of data and forwards it to the next unit, it need not sit idle until that set reaches its final destination in the logic path. Instead, it can immediately process the next set of data if it is made available. Wave-pipelining thus suggests simultaneous existence of multiple sets of data in a combinational logic circuit.

However, all the processing stages in a combinational circuit may not be identical, and different stages may incur different amounts of delay. Moreover, the data-dependent delays could cause more variations in the stage delays. Therefore, wave-pipelining for logic circuits involves significant design complexity and requires careful design and planning. Because of its susceptibility to the delay variations, wave-pipelining is used in relatively fewer logic circuits in practice.

On the other hand, the interconnects with repeaters are fairly regular circuits. Therefore, in the recent years, the focus of wave-pipelining has shifted from the logic circuits to the interconnect circuits [5], [16], and wave-pipelining is currently being investigated as a possible solution to achieve a significant enhancement in the interconnect throughput. By exploiting the techniques to minimize the variations in the

interconnect delay and/or designing timing and synchronization circuits that tolerate a certain amount of delay variation, wave-pipelining on the interconnects can become a very useful design technique for the high-performance global interconnects.

On the wave-pipelined interconnect circuits, a few interconnect segments can be considered to be one pipeline stage. Because of the structural regularity, all these stages are identical, which makes the design of the wave-pipelined interconnects relatively simple. In this technique, the repeater and wire capacitances are used to momentarily store the data bits and forward them to the next segment, and a new data bit can be sent on the interconnect before the previous data bit has reached its destination.

This high-speed serialization of the data leads to the idea of the communication throughput that is different from the reciprocal latency. On the wave-pipelined interconnect, because a new data bit is sent before the previous data bit reaches its destination, the maximum bit rate is not limited to the reciprocal latency. Instead, the minimum data pulsewidth that can be sustained on the wave-pipelined interconnect, which is smaller than the interconnect latency, determines the maximum interconnect throughput. Thus, wave-pipelining results in a significant enhancement in the communication throughput through simultaneous presence of multiple bits on the interconnect.

Figure 2.1 shows the impact of repeater insertion on the maximum bit rate that can be achieved on an interconnect. The comparison of bit rates obtained by the conventional latency-centric approach and the throughput-centric wave-pipelining approach is presented in this figure. As seen in Figure 2.1, for a large number of repeaters, throughput-centric repeater insertion significantly increases throughput because of the

simultaneous presence of multiple bits traveling along a single interconnect channel. However, the traditional reciprocal-latency-centric VLSI design has only a single bit on the channel at one time. As a result, the reciprocal latency predictions are very pessimistic compared to the actual bit rates that could be obtained in practice. For instance, it is seen from Figure 2.1 that the maximum bit rate predicted by the throughput-centric repeater insertion (4.44 Gbps) is more than seven times larger than the maximum reciprocal latency (0.63 Gbps).

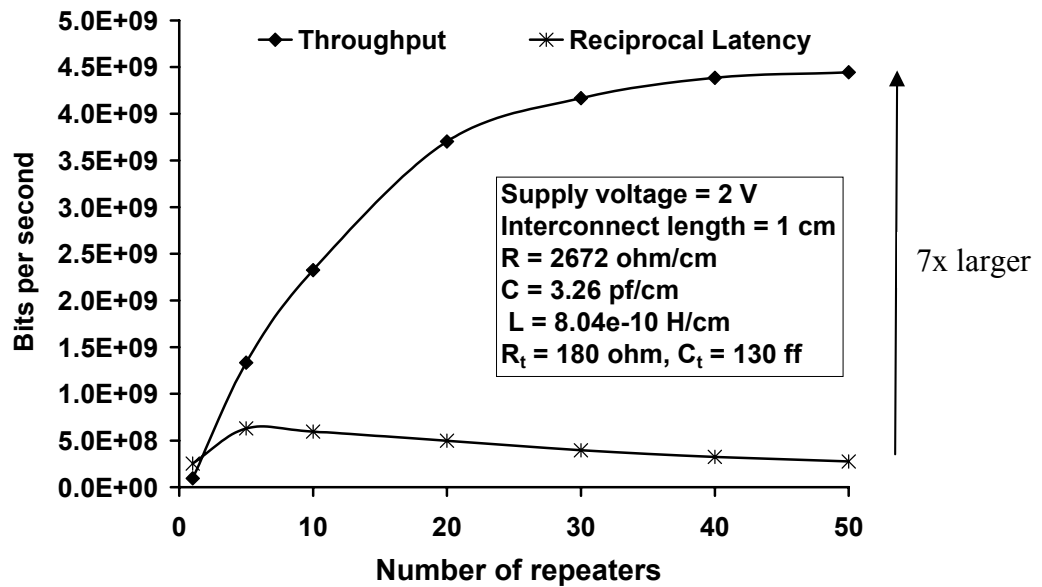


Figure 2.1: Difference between the values of throughput and reciprocal latency for a supply voltage of 2 V.

Therefore, the central thesis of this research is that the VLSI interconnect design strategies must shift from using global wires that can support only a single binary transition during the latency of the line to global wires that can sustain multiple bits traveling simultaneously along the length of the line. However, a large number of HSPICE simulations need to be run to calculate the throughput of such wave-pipelined interconnects, and this process can become tedious and time-consuming. Therefore, a

simple analytical expression to calculate the throughput of wave-pipelined interconnects is needed. The derivation of such an expression is presented in the following section.

2.3 Derivation and validation of analytical throughput model for wave-pipelined RC interconnects

2.3.1 Derivation of analytical throughput model

To ensure high-quality binary transmissions, it is assumed in the development of the analytical throughput model that the last repeated wire segment reaches at least 90% of the supply voltage. This voltage level can then reliably drive any latches or logic gates. This assumption regarding the 90% voltage swing is used in this section to maintain consistency with prior work in this area [6], [45]; however, this assumption is relaxed later in Section 2.5.1.

In the derivation of the closed-form throughput expression, a distributed RC model, as shown in Figure 2.2, is used to estimate the interconnect transients. Inductance is ignored based on [46], which states that the exclusion of inductance for a repeater-inserted interconnect gives negligible error in the results if the equivalent resistance of the interconnect segment is greater than its characteristic impedance. For an interconnect with repeaters, the equivalent resistance of the interconnect segment, R_{eq} , is given as

$$R_{eq} = R_{seg} + R_t, \quad (2.1)$$

where R_{seg} is the resistance of an interconnect segment and R_t is the output resistance of the repeater driver. The characteristic impedance for an interconnect segment can be defined as

$$Z_0' \approx \sqrt{\frac{L_{seg}}{C_{seg} + C_t}}, \quad (2.2)$$

where L_{seg} and C_{seg} are the inductance and capacitance of the interconnect segment, respectively, and C_t is the input capacitance of the repeater driver. (This expression for Z_0' is discussed in detail in Section 3.2.2.) The condition for ignoring inductance can be then written as

$$R_{seg} + R_t > \sqrt{\frac{L_{seg}}{C_{seg} + C_t}}. \quad (2.3)$$

It is seen in (2.3) that both R_{seg} and Z_0' decrease with the insertion of more repeaters, but R_t remains unchanged. Therefore, the inductance of a repeater-inserted interconnect can be ignored in the performance analysis if R_t is sufficiently larger than Z_0' .

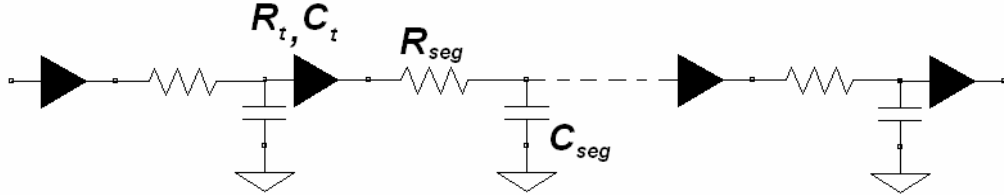


Figure 2.2: RC model for wave-pipelined interconnect circuit.

Figure 2.3 shows the cross-sectional dimensions for the interconnect. As shown in Figure 2.3, if w and h are the width and the height of the interconnect, respectively, s is separation between two interconnects, t is the thickness of the dielectric, and l is the length of the interconnect, the interconnect resistance R is given as

$$R = \frac{\rho l}{wh}, \quad (2.4)$$

where ρ stands for the resistivity of the interconnect metal. Because self inductance is ignored in this model, the skin effect is also not considered while calculating the interconnect resistance, to maintain consistency. However, ITRS provides the effective values of ρ for present and future technology generations, which can be used in (2.4) for first-order modeling of skin effect.

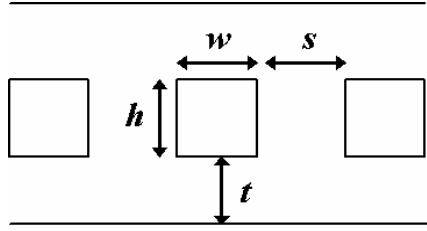


Figure 2.3: Interconnect cross-sectional dimensions.

To calculate the line capacitance, C , the interconnect is assumed to be placed between two co-planar interconnects and two orthogonal routing planes as shown in Figure 2.3. The capacitance per unit length is calculated using expressions in [36], and both parallel plate and fringing components have been considered. The driver output resistance, R_t , is calculated from the I-V curve of the MOSFET using HSPICE. The details of the estimation of R_t are discussed in Appendix A. The driver input capacitance, C_t , is calculated using expressions in [47] for level-49 transistor models.

The 1 cm long copper interconnect is assumed to be divided into n equal segments, each driven by a repeater. It is assumed that delay in each segment is given by the 50% rise time of the segment, and the delay of a repeater is denoted by $\Delta_{repeater}$. The variable v_k denotes the fraction of the supply voltage that is needed at the end of the k^{th} segment to achieve a 90% voltage swing at the output of the last segment.

The dominant pole approximation for the fraction of the voltage at the end of a line segment 'v' is given by [36] as

$$v = 1 - \sum_{i=1}^{\infty} K_i e^{-t_v/\sigma_i} \approx 1 - K_1 e^{-t_v/\sigma_1}. \quad (2.5)$$

Time t_v to reach v is then approximately given by

$$t_v = \sigma_{RCseg} \ln \left(\frac{K_1}{1-v} \right), \quad (2.6)$$

where $\sigma_{RCseg} = \sigma_1 = R_t C_t + R_t C_{seg} + C_t R_{seg} + 0.4 R_{seg} C_{seg}$ (2.7)

and $K_1 = 1.01 \left[\frac{R_t C_{seg} + R_{seg} C_t + R_{seg} C_{seg}}{R_t C_{seg} + R_{seg} C_t + \frac{\pi}{4} R_{seg} C_{seg}} \right]$. (2.8)

As in [45], t_k is the time required for the output of the k^{th} segment to reach a fraction v_k of the full scale supply voltage and is given by

$$t_k = \sigma_{RCseg} \ln \left(\frac{K_1}{1-v_k} \right) + (k-1) \sigma_{RCseg} \ln \left(\frac{K_1}{1-0.5} \right) + k \Delta_{repeater}, \quad (2.9)$$

where $\Delta_{repeater}$ is the 50% rise time of an inverter given by

$$\Delta_{repeater} = 0.693 R_t C_t. \quad (2.10)$$

Time t_{k-1} for the $(k-1)^{\text{th}}$ segment can be similarly written as

$$t_{k-1} = \sigma_{RCseg} \ln \left(\frac{K_1}{1-v_{k-1}} \right) + (k-2) \sigma_{RCseg} \ln \left(\frac{K_1}{1-0.5} \right) + (k-1) \Delta_{repeater}. \quad (2.11)$$

The transient analyses of various nodes of a repeater are shown in Figure 2.4. It is assumed that the inverter turns on when its input voltage reaches 50% of the supply voltage. Therefore, the output of an inverter is at its higher or lower peak when its input voltage is 50% of the supply voltage.

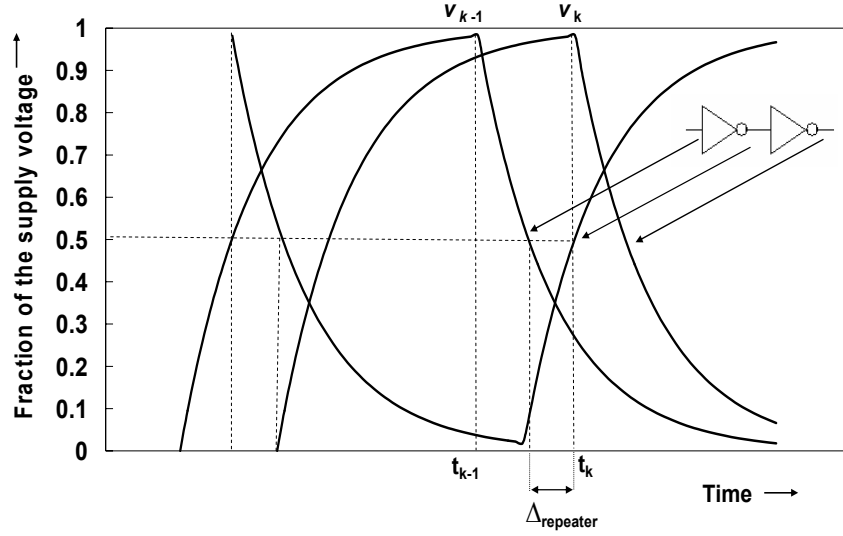


Figure 2.4: Waveforms at the input and output of k^{th} repeater.

It is seen in Figure 2.4 that the $(k-1)^{\text{th}}$ segment reaches a fraction v_{k-1} at time t_{k-1} , and in time $t_k - t_{k-1} - \Delta_{\text{repeater}}$, it discharges to 50% of the supply voltage. It can therefore be written as

$$v_{k-1} K_1 e^{-(t_k - t_{k-1} - \Delta_{\text{repeater}}) / \sigma_{RCseg}} = 0.5 . \quad (2.12)$$

Equation (2.12) reduces to

$$v_{k-1} K_1 e^{-\left(\ln 2 K_1 + \ln \left(\frac{1-v_{k-1}}{1-v_k} \right) \right)} = 0.5 , \quad (2.13)$$

which then leads to the recursive relationship

$$v_{k-1} = \frac{1}{2 - v_k} . \quad (2.14)$$

Therefore, for n repeater segments, (2.14) gives the output voltage swing of every repeater segment to achieve a v_n swing on the following segment. By setting v_n for the last segment to 0.9 (based on the assumption that the last segment reaches 90% of the supply voltage), v_1 for the first segment can be recursively calculated. Figure 2.5 shows

the value of v_n at the output of every repeated segment for an interconnect with 50 repeaters. *These values are independent of all technology attributes of the interconnect circuit and are fundamental to multibit transmission on VLSI interconnects with repeaters.*

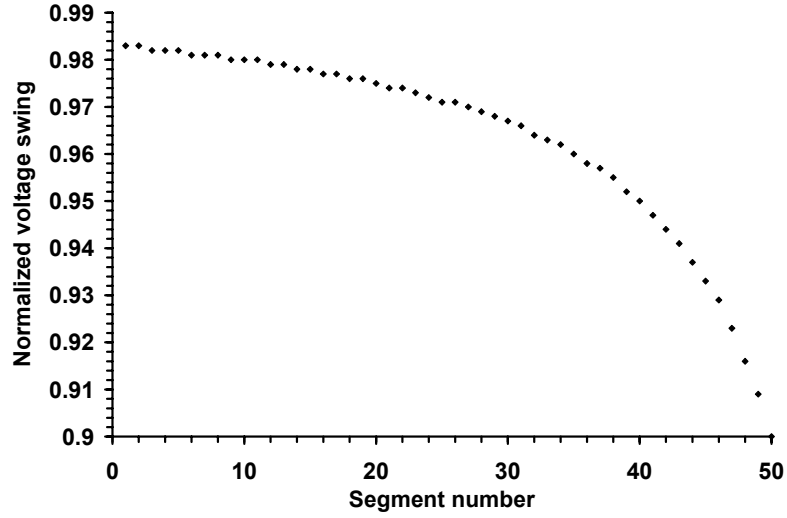


Figure 2.5: Values of v at the output of every repeated segment for an interconnect with 50 repeaters.

It is seen from Figure 2.5 that earlier segments need to undergo a swing of more than 90% to get a 90% swing at the output of the last segment. The first segment has to reach the highest value at the output compared to other segments. Therefore, the pulsewidth of the signal to be transmitted is limited by the first segment. The minimum pulsewidth, PW_{min} , is therefore given by

$$PW_{min} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater} . \quad (2.15)$$

The maximum throughput, T_{max} , is given by the inverse of the minimum pulsewidth as

$$T_{max} = \frac{1}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater}} . \quad (2.16)$$

The analytical throughput model is summarized in Table 2.1, whereas Table 2.2 is a direct look-up table for the values of v_1 . The left columns of Table 2.2 contain the total number of repeaters on the interconnect, and the right columns show the corresponding values of v_1 to be used in (2.16). The curve-fitting technique is used to get a simpler, non-recursive expression for the values of v_1 for up to 50 repeaters, which is shown in Figure 2.6. Figure 2.6 shows the values of v_1 using (2.14) and the expression obtained through curve-fitting. As seen in Figure 2.6, curve-fitting results in a less than 2% error in the values of v_1 for up to 50 repeaters.

Table 2.1: Summary of design equations for the analytical throughput model.

Parameter	Expression
Maximum throughput	$T_{\max} = \frac{1}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater}}$
Sakurai time constant [36]	$\sigma_{RCseg} = \sigma_1 = R_t C_t + R_t C_{seg} + C_t R_{seg} + 0.4 R_{seg} C_{seg}$
Sakurai coefficient [36]	$K_1 = 1.01 \left[\frac{R_t C_{seg} + R_{seg} C_t + R_{seg} C_{seg}}{R_t C_{seg} + R_{seg} C_t + \frac{\pi}{4} R_{seg} C_{seg}} \right]$
Voltage swing v_l at the output of the first repeated segment	$v_n = 0.9;$ For ($i=0; i < n; i++$) $v_{n-i} = 1/(2 - v_n);$ $v_l = v_{n-l};$
Internal time delay of a repeater	$\Delta_{repeater} = 0.693 R_t C_t$
Transistor equivalent resistance	R_t
Transistor equivalent capacitance	C_t
Interconnect segment resistance	R_{seg}
Interconnect segment capacitance	C_{seg}
Number of repeaters	n

Table 2.2: Values of v_1 for up to 50 repeaters.

n	v_1	n	v_1	n	v_1	n	v_1	n	v_1
1	0.900	11	0.950	21	0.967	31	0.975	41	0.980
2	0.909	12	0.952	22	0.968	32	0.976	42	0.980
3	0.916	13	0.955	23	0.969	33	0.976	43	0.981
4	0.923	14	0.957	24	0.970	34	0.977	44	0.981
5	0.929	15	0.958	25	0.971	35	0.977	45	0.981
6	0.933	16	0.960	26	0.971	36	0.978	46	0.982
7	0.937	17	0.962	27	0.972	37	0.978	47	0.982
8	0.941	18	0.963	28	0.973	38	0.979	48	0.982
9	0.944	19	0.964	29	0.974	39	0.979	49	0.983
10	0.947	20	0.966	30	0.974	40	0.980	50	0.983

n = Number of repeaters on the interconnect

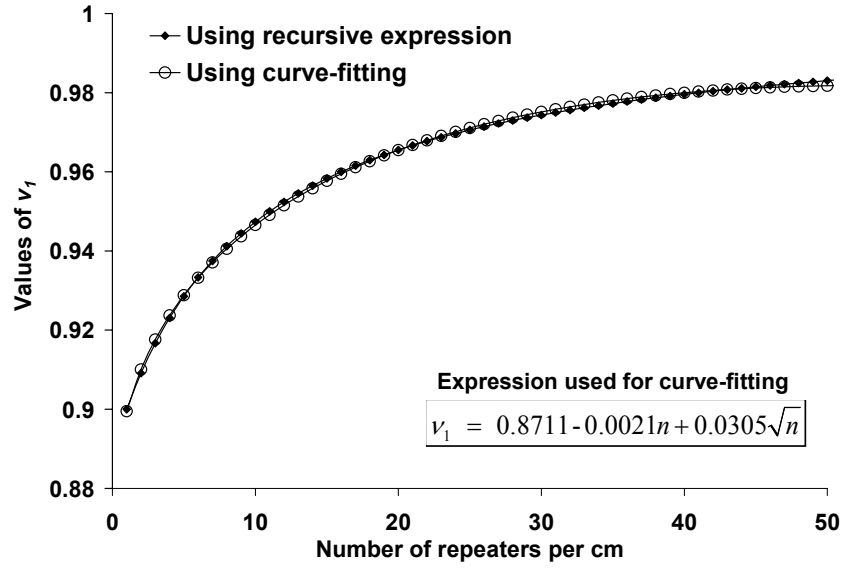


Figure 2.6: Curve-fitting for values of v_1 for up to 50 repeaters.

2.3.2 Validation of analytical throughput model

The values of throughput using the analytical model in (2.16) are compared with those from HSPICE level-49 models [48] for verification. The interconnect is 1 cm long and is modeled in HSPICE using a distributed resistance-inductance-capacitance (RLC)

network. It is assumed to drive a load equal to the input capacitance of a repeater. Various physical parameters used for the simulation are shown in Table 2.3. For the device and interconnect parameters in Table 2.3, the criterion in (2.3) is easily met. Therefore, the exclusion of inductance in the analytical model is justified.

Table 2.3: Physical parameters used for 180 nm HSPICE simulations.

Technology size	180 nm
Interconnect parameters ($w = h = s = t = 0.25 \mu\text{m}$)	
Resistance	2672 ohm/cm
Capacitance	3.26 pF/cm
Inductance	0.804 nH/cm
Repeater driver parameters (level-49 models)	
pMOS	Length = $0.18 \mu\text{m}$ and Width = $25.2 \mu\text{m}$
nMOS	Length = $0.18 \mu\text{m}$ and Width = $10.08 \mu\text{m}$

Figure 2.7 shows identical trends for the variation of throughput between the analytical model and HSPICE simulations for various supply voltages, with an average absolute error of 14%. The interconnect is represented by a dominant pole distributed RC model, which is a good approximation. However, simple driver models are used for repeaters, which cause the discrepancy between the analytical model and HSPICE simulation.

HSPICE results for voltages at different nodes of an interconnect circuit, whose dimensions are given in Table 2.3, are presented in Appendix B. These results are shown for a 1 cm long interconnect with 10 repeaters. It can be seen in Figure B.1 in Appendix B that there is no intersymbol interference (ISI) on the interconnect circuit and all the input data bits are correctly captured at the output. It can also be observed in Figure B.1 that the first repeated segment undergoes a larger voltage swing to achieve a 90% voltage swing at the output of the interconnect circuit.

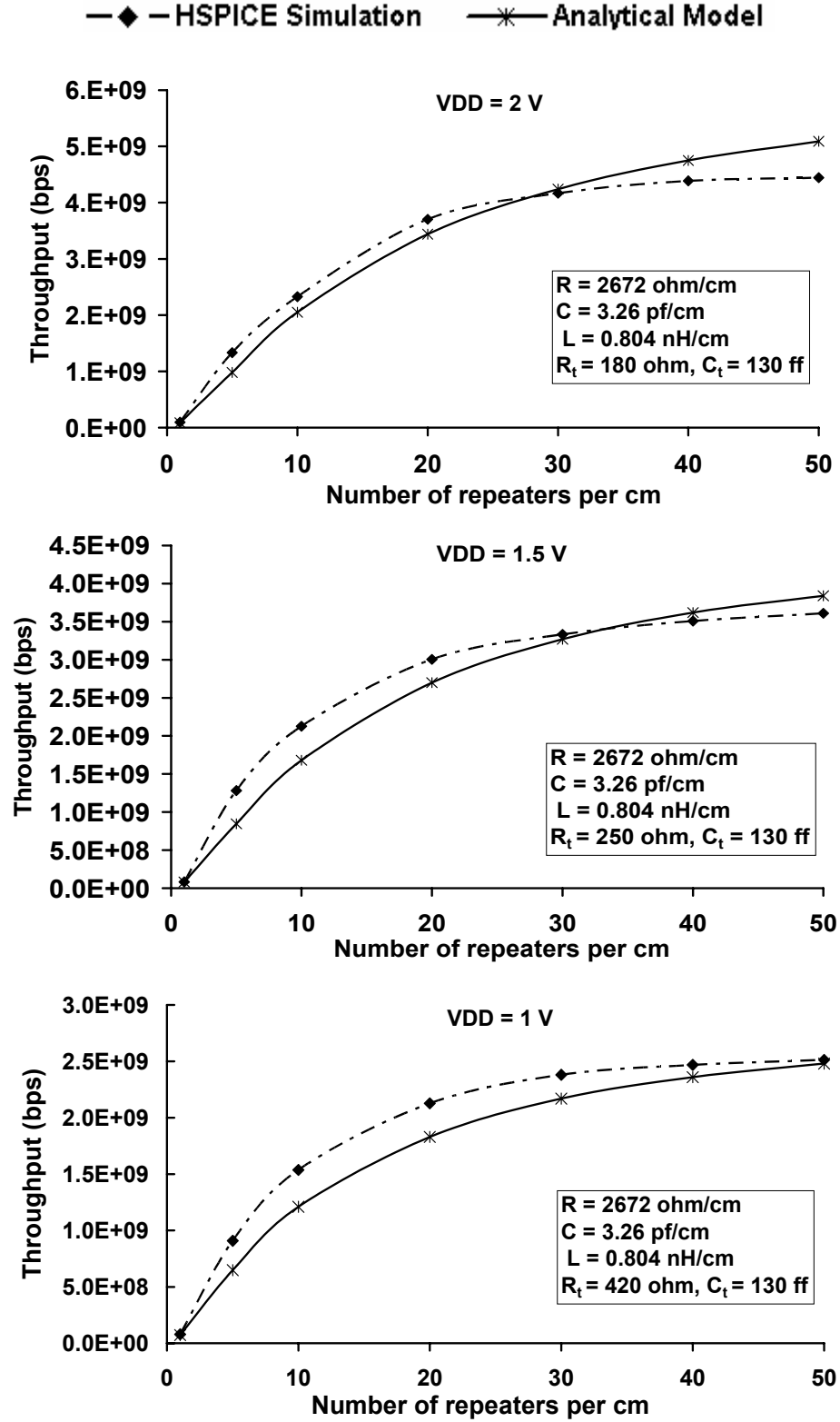


Figure 2.7: Comparison between results of closed-form analytical expression and HSPICE simulations.

It is interesting to compare the analytical throughput model in (2.16) with a more basic approximation based on the rise time of a repeated wire segment. If the bit rate is approximated by the inverse of the 90% rise time of a repeated wire segment,

$$\text{Bit rate} \approx \frac{1}{90\% \text{ rise time}} \approx \frac{1}{\sigma \ln\left(\frac{K_1}{1-0.9}\right)}. \quad (2.17)$$

The internal time delay of the repeater, Δ_{repeater} , is ignored in (2.17). The results for the bit rate using the analytical throughput model (ignoring Δ_{repeater}) and the expression in (2.17) are compared in Table 2.4 for the interconnect described by Table 2.3, for a 2 V supply. It is seen in Table 2.4 that both these approaches result in an identical bit rate for the single-driver interconnect because both the approaches suggest that the output of the interconnect needs to reach 90% of the supply voltage. However, as the number of repeaters increases, unlike (2.17), (2.16) suggests that the output of the first segment needs to reach more than 90% of the supply voltage, which results in a lower throughput. The comparison of the models given by (2.16) and (2.17) to HSPICE simulations in Figure 2.8 shows that (2.16) results in more accurate values of the bit rate compared to the simplistic approximation in (2.17).

Table 2.4: Comparison of bit rates using different models.

Number of repeaters	T_{\max} expression in (2.16) (bps)	Rise time expression in (2.17) (bps)
1	8.572E+07	8.572E+07
5	9.969E+08	1.161E+09
10	2.120E+09	2.722E+09
20	3.643E+09	5.323E+09
30	4.553E+09	7.239E+09
40	5.146E+09	8.682E+09
50	5.548E+09	9.800E+09

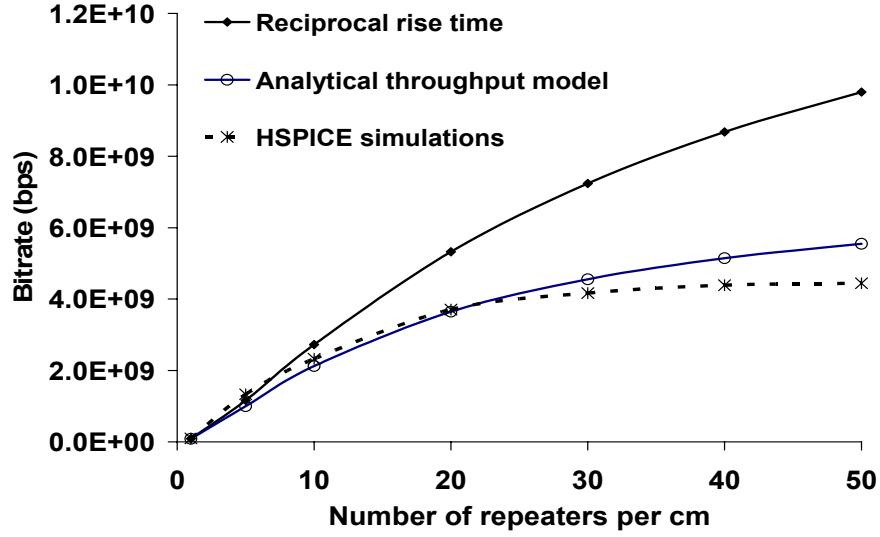


Figure 2.8: Comparison of bit rates using (2.16) and (2.17) with HSPICE simulations.

2.3.3 Multisegment coexistence of data

The minimum data pulsewidth that can be sustained on an interconnect is greater than the 50% rise time of a repeated wire segment. Consequently, even when a certain data bit triggers the repeater driver of the next segment, the following data bit does not immediately come on the previous repeated segment. Therefore, the same datum is present on multiple repeated interconnect segments at the same time. The number of segments, N_s , on which one data bit simultaneously exists, can be calculated as

$$N_s = \frac{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater}}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - 0.5} \right) + \Delta_{repeater}}, \quad (2.18)$$

where the term in the denominator stands for the 50% rise time of a repeated wire segment. The term $\Delta_{repeater}$ is significantly smaller than the first term in both the

numerator and the denominator of (2.18) in the normal operational region before saturation. Therefore, it can be neglected, and (2.18) can be rewritten as

$$N_s \approx \frac{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - \nu_1} \right)}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - 0.5} \right)}, \quad (2.19)$$

which further simplifies to

$$N_s \approx \ln \left(\frac{0.5}{1 - \nu_1} \right). \quad (2.20)$$

It is seen from (2.20) that like ν_l , the values of N_s are independent of all technology attributes of the interconnect circuit and are fundamental to multibit transmission on wave-pipelined VLSI interconnects. Using the values of ν_l in Table 2.2, the values of N_s are shown in Table 2.5 as the number of repeaters per cm (i.e., the repeater density) varies from 1 to 50. It is seen in Table 2.5 that as the repeater density increases, the number of segments on which the datum is simultaneously present also increases.

Table 2.5: Values of N_s for different repeater densities.

n	N_s	n	N_s	n	N_s	n	N_s	n	N_s
1	1.609	11	2.303	21	2.708	31	2.996	41	3.219
2	1.705	12	2.351	22	2.741	32	3.020	42	3.239
3	1.792	13	2.398	23	2.773	33	3.045	43	3.258
4	1.872	14	2.442	24	2.803	34	3.068	44	3.277
5	1.946	15	2.485	25	2.833	35	3.091	45	3.296
6	2.015	16	2.526	26	2.862	36	3.114	46	3.314
7	2.079	17	2.565	27	2.890	37	3.135	47	3.332
8	2.140	18	2.603	28	2.918	38	3.157	48	3.350
9	2.197	19	2.639	29	2.944	39	3.178	49	3.367
10	2.251	20	2.674	30	2.970	40	3.199	50	3.384

n = number of repeaters per cm

However, the segment length changes with the repeater density. Therefore, the values in Table 2.5 do not give any information about the effective interconnect length spanned by any data bit before the next data bit is sent on the interconnect. Table 2.6 shows the values for this effective bit length, l_{bit} , as the repeater density varies from 1 to 50. It is seen from Table 2.6 that though N_s increases with an increase in the repeater density, l_{bit} decreases, which translates into more data parallelism on the wave-pipelined interconnect for a larger number of repeaters.

Table 2.6: Values of l_{bit} for different repeater densities.

n	$l_{bit}(\text{cm})$	n	$l_{bit}(\text{cm})$	n	$l_{bit}(\text{cm})$	n	$l_{bit}(\text{cm})$	n	$l_{bit}(\text{cm})$
1	1.609	11	0.209	21	0.129	31	0.097	41	0.079
2	0.852	12	0.196	22	0.125	32	0.094	42	0.077
3	0.597	13	0.184	23	0.121	33	0.092	43	0.076
4	0.468	14	0.174	24	0.117	34	0.090	44	0.074
5	0.389	15	0.166	25	0.113	35	0.088	45	0.073
6	0.336	16	0.158	26	0.110	36	0.086	46	0.072
7	0.297	17	0.151	27	0.107	37	0.085	47	0.071
8	0.268	18	0.145	28	0.104	38	0.083	48	0.070
9	0.244	19	0.139	29	0.102	39	0.081	49	0.069
10	0.225	20	0.134	30	0.099	40	0.080	50	0.068

n = number of repeaters on the 1 cm long interconnect.

2.3.4 Saturation throughput

The analysis in Section 2.3.1 is extended in this subsection to derive the expression for the maximum saturation throughput that can be achieved on the wave-pipelined interconnect. Expanding σ_{RCseg} and $\Delta_{repeater}$ in terms of the device resistance and capacitance, (2.16) can be rewritten as

$$T_{\max} = \frac{1}{\left(R_t C_t + R_t C_{seg} + C_t R_{seg} + 0.4 R_{seg} C_{seg}\right) \ln\left(\frac{K_1}{1-v_1}\right) + 0.693 R_t C_t}. \quad (2.21)$$

If r and c are the interconnect resistance and capacitance per unit length, respectively, (2.21) can be further simplified as

$$T_{\max} = \frac{1}{\left(R_t C_t + R_t \frac{cl}{n} + C_t \frac{rl}{n} + 0.4 \frac{rcl^2}{n^2} \right) \ln \left(\frac{K_1}{1 - \nu_1} \right) + 0.693 R_t C_t}. \quad (2.22)$$

If h is the transistor scaling factor, R_t and C_t can be represented in terms of the resistance, R_0 , and capacitance, C_0 , of a minimum-sized repeater driver [6], respectively, as

$$R_t = \frac{R_0}{h} \text{ and } C_t = C_0 h. \quad (2.23)$$

Therefore, (2.22) can be further expanded to

$$T_{\max} = \frac{1}{\left(R_0 C_0 + \frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n} + 0.4 \frac{rcl^2}{n^2} \right) \ln \left(\frac{K_1}{1 - \nu_1} \right) + 0.693 R_0 C_0}. \quad (2.24)$$

An observation of Figure 2.7 shows that the communication throughput enters the saturation regime for a large value of n . It is then seen in the first term in the denominator of (2.24) that

$$0.4 \frac{rl}{n} \frac{cl}{n} \ll \frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n}. \quad (2.25)$$

Therefore, (2.24) can be rewritten as

$$T_{\max} \approx \frac{1}{\left[\ln \left(\frac{K_1}{1 - \nu_1} \right) + 0.693 \right] R_0 C_0 + \ln \left(\frac{K_1}{1 - \nu_1} \right) \left(\frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n} \right)}. \quad (2.26)$$

In the saturation regime, the transistor-dependent part of the pulsewidth equals its interconnect-dependent part, i.e.,

$$\left[\ln \left(\frac{K_1}{1-\nu_1} \right) + 0.693 \right] R_0 C_0 = \ln \left(\frac{K_1}{1-\nu_1} \right) \left(\frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n} \right). \quad (2.27)$$

In this region, as n becomes very large, both K_I and ν_I start to approach 1. Therefore, it is observed on the left hand side of the equation that

$$\ln \left(\frac{K_1}{1-\nu_1} \right) \gg 0.693. \quad (2.28)$$

Equation (2.27) then becomes

$$R_0 C_0 \approx \frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n}. \quad (2.29)$$

When the value of n/l obtained from (2.29) is substituted in (2.26) (with $K_I=1$), it leads to the expression of the saturation throughput as

$$T_{sat} \approx \frac{1}{\left[2 \ln \left(\frac{1}{1-\nu_1} \right) + 0.693 \right] R_0 C_0}. \quad (2.30)$$

Using (2.30), the saturation values of the throughput are presented in Table 2.7 for an average value of ν_I over 50 repeaters. A comparison of Table 2.7 with Figure 2.7 shows that the trend of variation of the saturation throughput with the supply voltage is captured well by (2.30).

Table 2.7: Saturation throughput using (2.30) for various supply voltages.

Supply voltage	Saturation throughput (Theory)	Saturation throughput (HSPICE)
2 V	5.82 Gbps	4.56 Gbps
1.5 V	4.19 Gbps	3.60 Gbps
1 V	2.49 Gbps	2.40 Gbps

As a side note, when the number of repeaters becomes asymptotically large, the analytical expression in (2.16) actually suggests a decrease in the throughput. Physically this occurs in the model because the first positive edge sent along the initially uncharged line propagates more slowly than the negative edge. This happens because the positive edge must rise through a larger voltage swing to reach the inverter threshold than does the falling edge. In this model, if the number of repeaters is large enough, any first signal pulse with a small pulsewidth sent along the line can be annihilated by such falling edge catching up to the rising edge. A similar behavior has been observed in HSPICE simulations as well. However, these instances occur well outside the normal range of repeater insertion and the exact mechanisms are outside the practical interest of this research. Equation (2.30) therefore does well on the onset of saturation where the most interesting design opportunities will occur.

To understand the effect of the technology-dependent parameters on the saturation throughput, simple first-order expressions for R_0 and C_0 are used. Assuming R_0 is approximately

$$R_0 \approx \frac{1}{\mu C_{ox} \left(\frac{W}{L} \right)_0 (V_{dd} - |V_t|)} , \quad (2.31)$$

where μ is the electron or hole mobility, C_{ox} is the oxide capacitance, $(W/L)_0$ is the aspect ratio of a minimum-sized transistor, V_{dd} is the supply voltage, and V_t is the threshold voltage [6], and

$$C_0 \approx C_{ox} L (W_{p0} + W_{n0}) , \quad (2.32)$$

where W_{p0} is the PMOS width and W_{n0} is the NMOS width for a minimum-sized MOSFET, (2.30) can be written as

$$T_{sat} \approx \frac{\mu(V_{dd} - |V_t|)}{2 \left[2 \ln \left(\frac{1}{1 - \nu_l} \right) + 0.693 \right] L^2}. \quad (2.33)$$

It is interesting to note from (2.33) that global wire scaling or transistor scaling does not affect the saturation throughput. For a given supply voltage, the saturation throughput depends only on the technology-dependent parameters. Though the expressions for the transistor parameters used here are first-order approximations, this analysis leads to the important conclusion that for a given supply voltage, *the technology generation alone sets the upper bound on the maximum communication throughput that can be obtained on an interconnect circuit.*

2.4 Impact of scaling on maximum wire bit rate

2.4.1 Length scaling

The dependence of T_{max} on the number of repeaters n and the interconnect length l can be seen in the expanded expression for the throughput as

$$T_{max} = \frac{1}{\left(R_t C_t + R_t \frac{cl}{n} + C_t \frac{rl}{n} + 0.4 \frac{rc l^2}{n^2} \right) \ln \left(\frac{K_l}{1 - \nu_l} \right) + 0.693 R_t C_t}. \quad (2.34)$$

Though ν_l also depends on the number of repeaters, it is seen in Table 2.2 that ν_l varies very slowly with respect to (w.r.t.) n , and its average value of 0.964 over 50 repeaters can be considered for simplicity. Under this assumption, (2.34) can be represented as a function of n/l ratio as

$$T_{\max} = \frac{1}{\left(R_t C_t + R_t \frac{c}{\left(\frac{n}{l} \right)} + C_t \frac{r}{\left(\frac{n}{l} \right)} + 0.4 \frac{rc}{\left(\frac{n}{l} \right)^2} \right) \ln \left(\frac{K_1}{1-v_1} \right) + 0.693 R_t C_t}. \quad (2.35)$$

The ratio n/l represents the number of repeaters per unit interconnect length. For a particular value of the throughput at a given supply voltage, only one value of this ratio satisfies (2.35). Therefore, *the first order of throughput approximately does not vary with the interconnect length or number of repeaters alone, but it is a function of their ratio.*

Though a 1 cm long interconnect is used for verification purposes in Section 2.3.2, (2.35) shows that similar results can be obtained for longer interconnects if the number of repeaters is increased in proportion. Figure 2.7 therefore shows the variation of throughput with the number of repeaters *per cm*. If v_l is assumed to be independent of n , the results obtained in the following chapters can be applicable not only to the specific case of a 1 cm long interconnect, but also any global interconnect of the 180 nm technology.

However, the impact of the underlying assumptions needs to be carefully analyzed to learn the scope and validity of this result. To make the length dependence analysis of the throughput more realistic, the assumption that v_l is independent of the repeater density is relaxed. Using the analytical throughput model in (2.35), Figure 2.9 shows the variation of throughput for various interconnect lengths as the repeater density changes from 10 to 50. These results are shown for the 180 nm interconnects having cross-sectional dimensions of 250 nm x 250 nm. A supply voltage of 2 V is used for these interconnects.

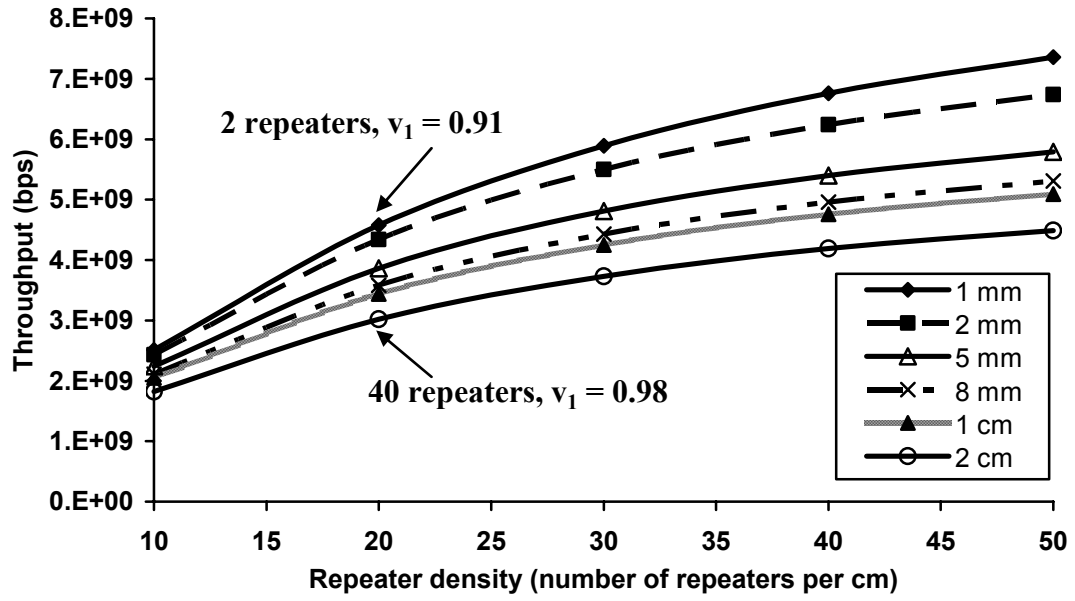


Figure 2.9: Variation of throughput with the repeater density for different interconnect lengths.

It is important to note that the x-axis in Figure 2.9 represents the repeater density, and not the absolute number of repeaters. For instance, a data point corresponding to 20 repeaters per cm translates into two repeaters for a 1 mm long interconnect and 40 repeaters for a 2 cm long interconnect. Though both these data points have the same repeater density, they have different numbers of repeaters inserted on them. As a result, they need to undergo different voltage swings at the output of the first segment to achieve a 90% voltage swing on the last segment. Table 2.2 shows that v_l equals 0.91 for two repeaters (1 mm interconnect) and it equals 0.98 for 40 repeaters (2 cm interconnect). This difference in v_l results in some difference in the maximum throughput that can be obtained for these two data points, as seen in Figure 2.9. Therefore, when a significant variation in the interconnect length is found in a particular set of interconnects, the throughput cannot be considered to be a function of the repeater density alone.

However, it is seen in Figure 2.9 that for larger interconnect lengths (8 mm – 2 cm), there is relatively less variation in the throughput for the same repeater density. This region consists of global interconnects, for which wave-pipelining using repeaters will be primarily used. Moreover, for a range of up to 15 repeaters per cm, which will be used in the most practical applications, the variation in the throughput for the same repeater density is further reduced. For instance, at 10 repeaters per cm, the maximum throughput given by an 8 mm long interconnect (that has eight repeaters) is almost equal to that given by a 1 cm long interconnect (that has ten repeaters). Therefore, in the region of practical interest, the throughput of global interconnects can be generalized to be a function of the repeater density alone, without a significant error in the results.

2.4.2 Constant field scaling

The impact of the constant field scaling on the interconnect throughput can also be studied with the help of the analytical expression in (2.22). If the constant field scaling is applied, the supply voltage and all the dimensions (including interconnect length) are scaled down by a factor of S ($S > 1$), and the throughput can be represented as

$$T_{\max} = \frac{1}{\frac{A}{S} + B}, \quad (2.36)$$

where

$$A = \left(R_t C_t + \frac{R_t c l}{n} \right) \ln \left(\frac{K_1}{1 - v_1} \right) + 0.693 R_t C_t \quad (2.37)$$

and

$$B = \left(\frac{r l C_t}{n} + \frac{0.4 r c l^2}{n^2} \right) \ln \left(\frac{K_1}{1 - v_1} \right). \quad (2.38)$$

It is observed for a single-driver interconnect that A is significantly smaller than B , which means that the throughput is almost unaffected by constant field scaling. However, as more repeaters are inserted, the term $0.4rc\ell^2/n^2$ in (2.38) begins to diminish because of the inverse quadratic dependence on n , which makes B much smaller than A , and the throughput increases approximately by the factor S . Constant field scaling thus results in a significant enhancement in the throughput for a large number of repeaters. Therefore, it can be concluded that the constant field scaling improves the maximum saturation throughput by a factor S .

2.4.3 Interconnect scaling

For the results shown in Figure 2.7, a square cross-section of 250 nm width is assumed for the global interconnect. However, if this interconnect is placed on the less densely packed upper metal layers, a higher cross-sectional width such as 1000 nm could be used. The interconnect capacitance is a function of the ratios of the cross-sectional dimensions w , h , s , and t [36]. Even when a cross-sectional width of 1000 nm is used, if $w = h = s = t$, the interconnect capacitance is unchanged. However, the interconnect resistance varies inversely with the square of the interconnect width and becomes 16 times smaller. Therefore, it can be deduced from (2.21) that there will be an increase in the throughput, which is also shown in Figure 2.10. However, it is important to note that this increase occurs at the expense of a four times increase in the wire area, and the maximum saturation throughput, which is shown by (2.30) to be a function of only transistor parameters, does not change. Similarly, if a smaller width of 180 nm is used for

the square cross-section, the throughput decreases as shown in Figure 2.10, along with a reduction in the wire area.

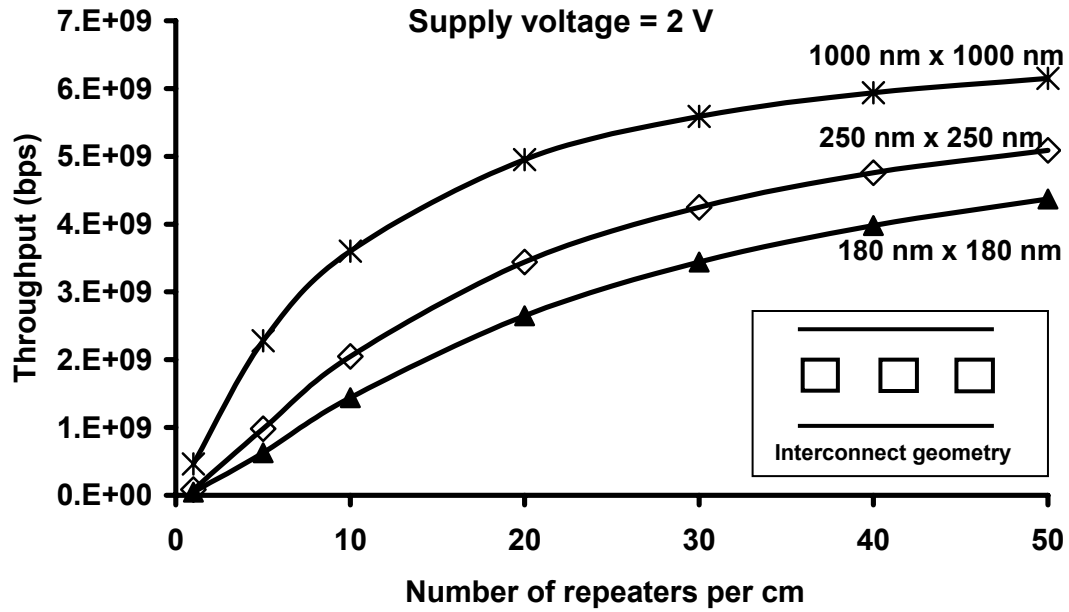


Figure 2.10: Variation of throughput with number of repeaters for different square cross-sections for a 2 V supply, using analytical throughput model.

The wire-sizing opportunities to enhance throughput when the wire area is constrained can be studied by observing the variation of throughput w.r.t. the wire width for a constant wire pitch. Therefore, Figure 2.11 shows the variation of throughput with the interconnect width for different interconnect aspect ratios (AR) and a constant wire pitch of 1000 nm. It is seen in Figure 2.11 that as the wire dimensions initially become larger, the interconnect resistance decreases, which results in an increase in the throughput.

However, for a constant pitch, as the interconnect width increases, the spacing between two interconnects decreases. Therefore, the increase in the wire width for a constant pitch results in some increase in the ground capacitance of the interconnect and a significant increase in its mutual capacitance. Beyond a certain value of the interconnect width, this increase in the interconnect capacitance overshadows the decrease in the interconnect resistance, and the throughput starts to decrease.

Therefore, for a constant interconnect AR and fixed values of the pitch and the dielectric thickness, there is an optimal value of the interconnect width that maximizes the throughput. It can also be deduced from Figure 2.11 that for smaller interconnect widths, a higher AR results in a higher throughput because of the reduced interconnect resistance. However, for larger interconnect widths, an increase in AR significantly increases the mutual capacitance, which becomes a more dominant factor, thereby reducing the throughput. It should also be noted from Figure 2.11 that the values of maximum throughput achieved on this interconnect are identical for all values of AR.

It is seen from Figure 2.11 that the lowest value of the interconnect width that achieves the maximum throughput is approximately $0.25\text{ }\mu\text{m}$ ($\text{AR} = 2$). For a constant pitch, operating with this interconnect width translates into a lower interconnect capacitance, which reduces power. Operating with a smaller width also translates into a larger spacing, which could reduce crosstalk. Therefore, Figure 2.11 shows that it is not necessary to use large interconnect widths to achieve high throughput. Smaller interconnect widths, along with a high aspect ratio, can not only achieve high throughput but also reduce power and crosstalk compared to those for larger interconnect widths.

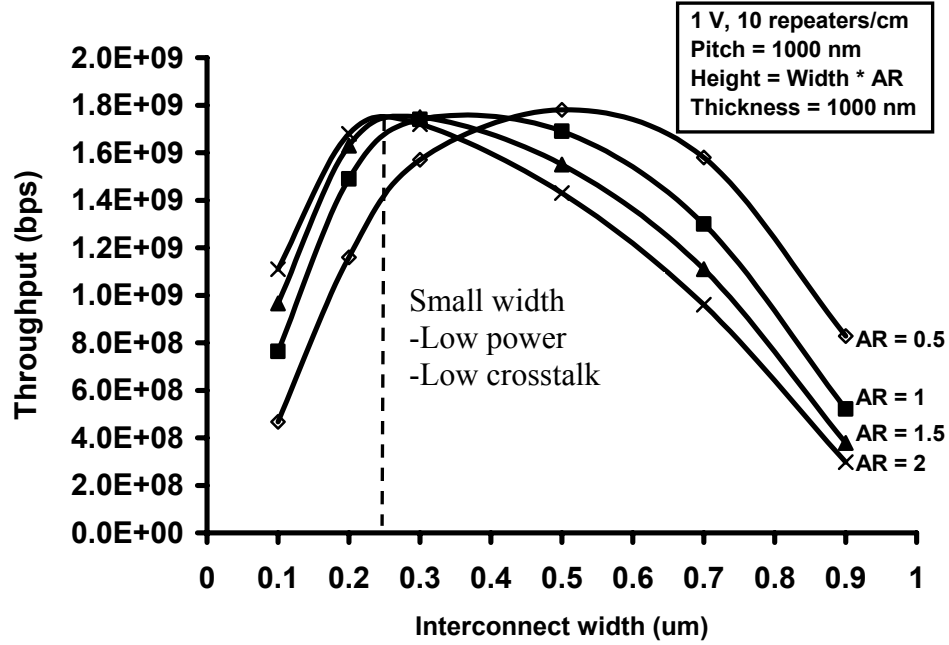


Figure 2.11: Variation of throughput with interconnect width for constant pitch for different interconnect aspect ratios, using (2.16).

2.4.4 Transistor width scaling

If the minimum-sized transistors are used to design the repeater circuits, the corresponding throughput can be written as

$$T_{\max_0} = \frac{1}{\left(R_0 C_0 + R_0 C_{seg} + C_0 R_{seg} + 0.4 R_{seg} C_{seg} \right) \ln \left(\frac{K_1}{1 - \nu_1} \right) + 0.693 R_0 C_0} \cdot (2.39)$$

However, if the transistor width is scaled up by a factor h ($h > 1$), R_0 changes to R_0/h and C_0 changes to $C_0 h$. Equation (2.39) can then be rewritten as

$$T_{\max} = \frac{1}{\left[\ln \left(\frac{K_1}{1 - \nu_1} \right) + 0.693 \right] R_0 C_0 + \ln \left(\frac{K_1}{1 - \nu_1} \right) \left(0.4 \frac{rl}{n} \frac{cl}{n} + \frac{R_0}{h} \frac{cl}{n} + C_0 h \frac{rl}{n} \right)} \cdot (2.40)$$

It is seen from (2.40) that transistor scaling creates two competing terms in the denominator, one varying directly with h and the other varying inversely with it. Therefore, it is necessary to choose h such that it minimizes the total sum in the denominator, thereby giving the highest throughput. Partially differentiating (2.40) w.r.t. h and setting the result equal to zero, the optimal value of h is found to be

$$h_{opt} = \sqrt{\frac{R_0 c}{C_0 r}}, \quad (2.41)$$

which is same as that obtained in [6]. It is interesting to note that though [6] proposes latency-centric repeater insertion and this research proposes throughput-centric repeater insertion, both these works result in the same expression for the optimal repeater scaling factor. It should also be noted that h_{opt} is independent of the number of repeaters.

A quasi-optimal transistor scaling factor can be obtained for this throughput-centric design to reduce silicon area and power at the expense of a small reduction in the throughput. Assuming that the contribution of transistor-dependent parameters to pulsewidth is same as that of interconnect-dependent parameters in the denominator of (2.40), the quasi-optimal scaling factor is given as

$$h_{quasi,opt} = \frac{R_0}{2 R_{seg}} \pm \sqrt{\left(\frac{R_0}{2 R_{seg}}\right)^2 - h_{opt}^2}, \quad (2.42)$$

For an interconnect with resistance of 172 ohm/cm and capacitance of 2.226 pF/cm, the value of the optimal transistor scaling factor is 237 and that of the quasi-optimal scaling factor is 121 when 10 repeaters are inserted per unit cm length. The use of the quasi-optimal factor in this case results in a 10% decrease in throughput compared to that with the optimal scaling factor, but it reduces the silicon area by almost 50%. However, it is

seen in (2.42) that $h_{quasi,opt}$ depends on the repeater density and is therefore different for different designs. Therefore, the optimal scaling factor h_{opt} is used in this analysis for its simplicity and generality.

For a large number of repeaters, the throughput obtained by using the minimum-sized repeaters is comparable to that using the optimal-sized repeaters because the second term in the denominator of (2.40) becomes very small. This result is also supported by the expression of the saturation throughput in (2.30), which is independent of h . However, for a moderate number of repeaters, the use of the optimal-sized repeaters results in a significantly higher throughput than that given by the minimum-sized repeaters.

The presence of R_0 in (2.41) shows that h_{opt} depends on the supply voltage. For the validation results shown in Figure 2.7, an average value of h_{opt} over the chosen supply voltage range is found to be 56. Therefore, the NMOS and PMOS aspect ratios are chosen to be 56 and 140, respectively. The ratio of NMOS width to PMOS width is chosen to be 1:2.5 for equal rise and fall time design.

Table 2.8 presents a summary of Section 2.4 by showing the impact of scaling on maximum throughput. The scaling factors for certain transistor and interconnect parameters and the corresponding scaling factors for the throughput are shown in Table 2.8. The regions in which these generalized scaling properties are applicable are also shown in Table 2.8.

Table 2.8: Summary of impact of scaling on maximum throughput.

Type of scaling	Scaled parameter and scaling factors	Throughput scaling factor	Region where this generalization is applicable
Length scaling	$(n/l) \rightarrow 1/S$	1	Global interconnects, constant and moderate repeater density
Constant field scaling	$V_{dd} \rightarrow 1/S$, all dimensions $\rightarrow 1/S$	S	Saturation region
Interconnect scaling	$w, h, s, t \rightarrow S$	$< S^2$	Square cross-sections
Transistor width scaling	$W_n, W_p \rightarrow S$	Between $1/S$ and S	Throughput increases up to h_{opt} , decreases thereafter

2.5 Impact of changes in design assumptions on throughput

The analytical throughput expression in (2.16) is based on certain assumptions about the construction and working of repeater circuits. However, it would be interesting to see the impact of changes in some of these design assumptions on the communication throughput. Therefore, the analytical throughput model is re-derived for each of the following cases.

1. The last segment is required to achieve a normalized voltage swing other than 0.9.
2. The inverters in the repeater have a normalized switching threshold other than 0.5.
3. A repeater consists of a single inverter instead of two.

The relevant equations in the derivation of the analytical throughput expression in Section 2.3.1 are rewritten in this section for convenience. In these equations, R_t and C_t denote the resistance and capacitance of a transistor, and R_{seg} and C_{seg} denote the resistance and capacitance of an interconnect segment, respectively.

$$\sigma_{RCseg} = R_t C_t + R_t C_{seg} + C_t R_{seg} + 0.4 R_{seg} C_{seg} . \quad (2.43)$$

$$K_1 = 1.01 \left[\frac{R_t C_{seg} + R_{seg} C_t + R_{seg} C_{seg}}{R_t C_{seg} + R_{seg} C_t + \frac{\pi}{4} R_{seg} C_{seg}} \right] . \quad (2.44)$$

$$\Delta_{repeater} = 0.693 R_t C_t . \quad (2.45)$$

$$t_k = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_k} \right) + (k-1) \sigma_{RCseg} \ln \left(\frac{K_1}{1 - 0.5} \right) + k \Delta_{repeater} . \quad (2.46)$$

$$t_{k-1} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{k-1}} \right) + (k-2) \sigma_{RCseg} \ln \left(\frac{K_1}{1 - 0.5} \right) + (k-1) \Delta_{repeater} . \quad (2.47)$$

$$v_{k-1} K_1 e^{-(t_k - t_{k-1} - \Delta_{repeater}) / \sigma_{RCseg}} = 0.5 . \quad (2.48)$$

$$v_{k-1} = \frac{1}{2 - v_k} . \quad (2.49)$$

$$PW_{\min} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater} . \quad (2.50)$$

$$T_{\max} = \frac{1}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_1} \right) + \Delta_{repeater}} . \quad (2.51)$$

2.5.1 The last repeated wire segment achieves a normalized voltage swing other than 0.9

The analytical throughput model assumes that the last repeated wire segment reaches $0.9V_{dd}$ (i.e., $v = 0.9$ on the last segment) for simplicity and consistency with prior work in this area. However, for the successful latching of data, the data bit at the output

needs to be stable for the time equal to the hold time of the latching circuitry in addition to the 50% rise time, which could be less than the 90% rise time. Therefore, there is an opportunity to send the data bits at a faster rate than that achieved with the initial assumption about the 90% rise time.

To capture such effects, the analytical throughput model can be generalized for any value of v at the output. Despite this generalization, all the equations in the derivation of the throughput model, including the final expression (2.51), remain unchanged. The only parameter that changes with this generalization is the value of v_I , i.e., the normalized voltage swing at the output of the first segment.

The recursive values of v are shown in Table 2.9 over 20 iterations of (2.49). These values correspond to the normalized voltage swings of 0.5, 0.85, 0.95, and 0.99 at the output of the last segment (which are the values corresponding to Iteration 1). The values of v corresponding to the normalized last segment voltage swing of 0.9, which is the base case, are also included in Table 2.9 for comparison.

The last row of Table 2.9 shows the values of v_I corresponding to different values of the normalized voltage swing at the output of the last segment, for an interconnect with 20 repeaters. For instance, to achieve a $0.5V_{dd}$ voltage swing at the output of the last segment on an interconnect with 20 repeaters, the first segment needs to undergo a voltage swing of $0.952V_{dd}$. The results in Table 2.9 suggest that to achieve a large voltage swing at the output of the last repeated wire segment in a wave-pipelined interconnect, the first segment must undergo a proportionally larger voltage swing. Conversely, if a normalized voltage swing smaller than 0.9 is sufficient to trigger the output circuits that

the interconnect is driving, a smaller v_I is required, which in turn results in a larger communication throughput.

Table 2.9: Values of v for different normalized voltage swings at the output of the last segment (over 20 iterations).

Iteration	Values of v				
1	0.900	0.500	0.850	0.950	0.990
2	0.909	0.667	0.870	0.952	0.990
3	0.917	0.750	0.885	0.955	0.990
4	0.923	0.800	0.897	0.957	0.990
5	0.929	0.833	0.906	0.958	0.990
6	0.933	0.857	0.914	0.960	0.990
7	0.938	0.875	0.921	0.962	0.991
8	0.941	0.889	0.927	0.963	0.991
9	0.944	0.900	0.932	0.964	0.991
10	0.947	0.909	0.936	0.966	0.991
11	0.950	0.917	0.940	0.967	0.991
12	0.952	0.923	0.943	0.968	0.991
13	0.955	0.929	0.946	0.969	0.991
14	0.957	0.933	0.949	0.970	0.991
15	0.958	0.938	0.952	0.971	0.991
16	0.960	0.941	0.954	0.971	0.991
17	0.962	0.944	0.956	0.972	0.991
18	0.963	0.947	0.958	0.973	0.991
19	0.964	0.950	0.959	0.974	0.992
20	0.966	0.952	0.961	0.974	0.992

2.5.2 The inverters in a repeater have a normalized switching threshold other than 0.5

The analytical throughput expression in (2.51) also assumes that the inverter in a repeater transitions to a different state when its input voltage is $0.5V_{dd}$. However, the CMOS circuit theory defines an inverter threshold, V_I , which is commonly used as the switching threshold voltage for an inverter. V_I is defined by the point at which the voltage

transfer curve of the inverter intersects with the unity gain line [49]. The voltage V_I is given by

$$V_I = \frac{V_{dd} - |V_{tp}| + \sqrt{\frac{\beta_n}{\beta_p}} V_{tn}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}, \quad (2.52)$$

where V_{tn} is the NMOS threshold voltage and V_{tp} is the PMOS threshold voltage. The terms β_n and β_p are given by

$$\beta_n = \mu_n C_{ox} \left(\frac{W}{L} \right)_n \quad (2.53)$$

and

$$\beta_p = \mu_p C_{ox} \left(\frac{W}{L} \right)_p, \quad (2.54)$$

where μ_n and μ_p are the mobilities of electrons and holes, respectively, and C_{ox} is the capacitance of the gate oxide. The terms $(W/L)_n$ and $(W/L)_p$ denote the aspect ratios of the NMOS and PMOS, respectively.

The values of V_I for two supply voltages are shown in Table 2.10 for an inverter scaling factor of 56. These values are calculated using the 180 nm MOSIS parameters [48]. It is seen in Table 2.10 that V_I is very close to $0.5V_{dd}$, and the analytical throughput model in (2.51) is based on a realistic assumption.

Table 2.10: Values of inverter threshold voltage for two different supply voltages.

Supply voltage V_{dd}	2 V	1 V
Inverter threshold voltage V_I	0.993 V	0.481 V
V_I / V_{dd}	0.497	0.481

Assuming v_I is the inverter threshold voltage normalized to the supply voltage, (2.46) and (2.47) can be rewritten as

$$t_k = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_k} \right) + (k-1) \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_I} \right) + k \Delta_{repeater} \quad (2.55)$$

and
$$t_{k-1} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{k-1}} \right) + (k-2) \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_I} \right) + (k-1) \Delta_{repeater} . \quad (2.56)$$

Equation (2.48) becomes

$$v_{k-1} K_1 e^{-(t_k - t_{k-1} - \Delta_{repeater}) / \sigma_{RCseg}} = v_I , \quad (2.57)$$

which then leads to a recursive relationship as

$$v_{k-1} = \frac{\frac{v_I}{1 - v_I}}{\left(\frac{v_I}{1 - v_I} + 1 \right) - v_k} . \quad (2.58)$$

If
$$C' = \frac{v_I}{1 - v_I} , \quad (2.59)$$

(2.58) can be rewritten as
$$v_{k-1} = \frac{C'}{(C' + 1) - v_k} . \quad (2.60)$$

Beyond this point, the expressions for the pulsewidth and the throughput remain unchanged.

If $v_I = 0.5$ (as in the initial assumption), $C' = 1$ and (2.49) becomes a special case of the relationship given in (2.60). Table 2.11 shows the normalized voltage swings at the output of the repeated interconnect segments using 20 iterations of (2.49), for the values of v_I shown in Table 2.10, along with the base case v_I of 0.5. It is seen in Table 2.11 that smaller values of v_I result in the smaller values of v at the output of the first segment (v_I),

which in turn increases the throughput. Therefore, it may be useful to size the repeaters such that they switch at $v_I < 0.5$, which would then result in some increase in the throughput performance.

Table 2.11: Values of v for different inverter threshold voltages (over 20 iterations).

Iteration	$v_I = 0.5$	$v_I = 0.497$	$v_I = 0.481$
1	0.900	0.900	0.900
2	0.909	0.908	0.903
3	0.917	0.915	0.905
4	0.923	0.921	0.907
5	0.929	0.926	0.909
6	0.933	0.930	0.910
7	0.938	0.934	0.912
8	0.941	0.937	0.913
9	0.944	0.940	0.914
10	0.947	0.943	0.915
11	0.950	0.945	0.916
12	0.952	0.948	0.917
13	0.955	0.950	0.918
14	0.957	0.952	0.919
15	0.958	0.953	0.919
16	0.960	0.955	0.920
17	0.962	0.956	0.921
18	0.963	0.958	0.921
19	0.964	0.959	0.922
20	0.966	0.960	0.922

2.5.3 A repeater consists of a single inverter instead of two

The analytical throughput model in (2.51) is based on the assumption that a repeater consists of two inverters. The repeaters used in the derivation of throughput are thus non-inverting. However, a single inverter can also be used as a repeater as long as an even number of inverting repeaters are inserted on the interconnect to ensure the correct

data polarity at the output. The modification in (2.51) for a single inverter repeater is presented in this subsection.

Figure 2.12 shows an interconnect with the inverting repeaters. The normalized voltage swing at every repeated segment is subscripted by ‘o’ (for odd) or ‘e’ (for even), based on the placement number of that repeater. For instance, the normalized voltage swing corresponding to the first inverting repeater is subscripted by ‘o’ because 1 is an odd number and that corresponding to the second inverting repeater is subscripted by ‘e’ because 2 is an even number.

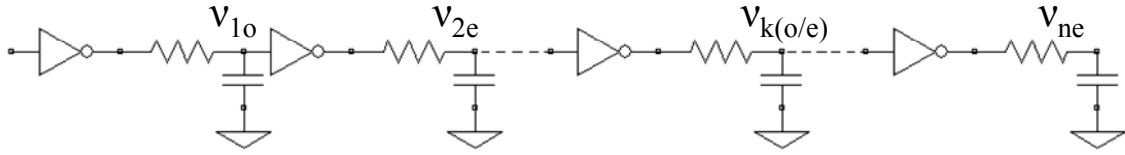


Figure 2.12: An RC interconnect with inverting repeaters.

If the input datum switches from 0 to 1, the output of the first segment and every odd-placed segment goes from 1 to 0. The switching time required for the output of any k^{th} odd repeated segment is given by

$$t_{ko} = \sigma_{RCseg} \ln \left(\frac{K_1}{v_{ko}} \right) + (k-1) \sigma_{RCseg} \ln \left(\frac{K_1}{1-0.5} \right). \quad (2.61)$$

The terms v with a subscript ‘o’ will be smaller than 0.5 because they correspond to a discharging event. The term v_{ko} can be redefined in terms of v_{ke} as

$$v_{ko} = 1 - v_{ke}, \quad (2.62)$$

where v_{ke} is greater than 0.5. Equation (2.46) can then be rewritten as

$$t_{ko} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{ke}} \right) + (k-1) \sigma_{RCseg} \ln \left(\frac{K_1}{1-0.5} \right). \quad (2.63)$$

The $(k-1)^{\text{th}}$ segment corresponds to an even-placed repeater, therefore,

$$t_{(k-1)e} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{(k-1)e}} \right) + (k-2)\sigma_{RCseg} \ln \left(\frac{K_1}{1 - 0.5} \right). \quad (2.64)$$

From (2.63) and (2.64), it can be written as

$$v_{(k-1)e} K_1 e^{-(t_{ko} - t_{(k-1)e}) / \sigma_{RCseg}} = 0.5, \quad (2.65)$$

which is similar to (2.48). Therefore, the analysis beyond this point is unchanged. The only difference between (2.61)-(2.65) and the original equations (2.46)-(2.48) is the absence of the term $\Delta_{repeater}$. When the repeater consists of only a single inverter, the internal time delay of the repeater does not need to be accounted for. Therefore, the pulsewidth is given by

$$PW_{\min} = \sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{1e}} \right), \quad (2.66)$$

and the throughput is given by

$$T_{\max} = \frac{1}{\sigma_{RCseg} \ln \left(\frac{K_1}{1 - v_{1e}} \right)}. \quad (2.67)$$

The v terms with a subscript ‘e’ are identical to the v terms in the original derivation of the throughput model for non-inverting repeaters. Therefore, the values of v in Table 2.2 can be directly used for the case of inverting repeaters.

The comparison of (2.67) and (2.51) shows that for the same number of repeaters, the maximum throughput that can be achieved on an interconnect with inverting repeaters is higher than that for an interconnect with non-inverting repeaters. This fact is also reflected in the HSPICE simulation results shown in Table 2.12 for a 180 nm metal-5

interconnect in [29], whose dimensions are shown in Table 1.1. The absence of the term $\Delta_{repeater}$ in the case of inverting repeaters allows data transmission using a smaller pulsewidth compared to that with non-inverting repeaters, thereby enhancing the maximum communication throughput.

Table 2.12: Comparison of throughput for non-inverting and inverting repeaters using HSPICE simulations.

Number of repeaters per cm	Throughput (bps) with non-inverting repeaters	Throughput (bps) with inverting repeaters
2	1.444E+09	1.478E+09
6	3.505E+09	3.716E+09
10	4.555E+09	4.919E+09
20	5.870E+09	6.487E+09
30	6.480E+09	7.241E+09
40	6.666E+09	7.474E+09
50	6.816E+09	7.662E+09

2.6 Summary

The importance of wave-pipelining using repeaters is discussed in this chapter to enhance the throughput performance of VLSI global interconnect circuits. A closed-form analytical expression to calculate the throughput of wave-pipelined RC interconnects is derived in this chapter and is successfully validated using HSPICE simulations for RLC interconnect circuits. One of the most important applications of the analytical throughput model is its use to study the limits and opportunities of wave-pipelining in VLSI global interconnects. Therefore, the analytical throughput model is used to analyze the effect of various transistor and interconnect parameters on the interconnect throughput and also calculate the maximum saturation throughput that can be achieved on an interconnect.

CHAPTER 3

SIGNAL INTEGRITY ANALYSIS OF WAVE-PIPELINED INTERCONNECTS

3.1 Introduction

The analytical throughput model for the wave-pipelined RC interconnects is presented in the previous chapter, and the impact of inductance on throughput and latency of the interconnect is analyzed in this chapter. Different performance models based on bandwidth, delay, and rise time are discussed in this chapter for the low-loss transmission line. Based on [46], the boundary between RC and RLC characterization of the interconnect circuits is also discussed in this chapter.

To analyze the impact of wave-pipelining on signal integrity in the presence of dynamic delay effects resulting from inductive and capacitive coupling, a distributed RLC interconnect system is modeled using RAPHAEL and HSPICE. The impact of wave-pipelining on the throughput, latency, overshoot voltage, and crosstalk voltage is analyzed with the help of HSPICE simulations. Finally, techniques such as inserting shielding ground lines, inserting misaligned repeaters, and inserting decoupling capacitors are discussed to dampen the effect of inductive and capacitive coupling on the interconnect performance and signal integrity (e.g., dynamic delay effects).

3.2 Performance models for transmission line

The theoretical analysis of the interconnect throughput in the previous chapter has accounted for the interconnect resistance, R , and the interconnect capacitance, C . However, the interconnect also has a finite inductance, L , associated with it. To understand the impact of inductance on the interconnect performance, different performance models for RLC interconnects are discussed in this chapter. This section discusses these performance models for the low-loss VLSI transmission line.

The interconnect dimensions and the corresponding parasitic values for the 180 nm transmission line used in this chapter are shown in Table 3.1. In Table 3.1, K_m and K_{LR} are the inductance coupling factors, which are given by the ratio of the mutual inductance to the self inductance. The transmission line is also represented graphically in Figure 3.1 for clarity.

Table 3.1: Dimensions and parasitics for transmission line.

Supply voltage	$V_{dd} = 2 \text{ V}$
Interconnect dimensions	$l = 1 \text{ cm}, w = s = 1 \text{ } \mu\text{m}, h = t = 2.5 \text{ } \mu\text{m}$
Interconnect resistance	$R = 88 \text{ ohm/cm}$
Interconnect capacitance	$2C_g = 0.5 \text{ pF/cm}, 2C_m = 1 \text{ pF/cm}$
Interconnect self inductance and coupling factors	$L = 3.6 \text{ nH/cm}$, Near neighbors $K_m = 0.5738$, Far neighbors $K_{LR} = 0.3317$
Input resistance and load capacitance	$R_{in} = 180 \text{ ohm}$, $C_L = 100 \text{ fF}$

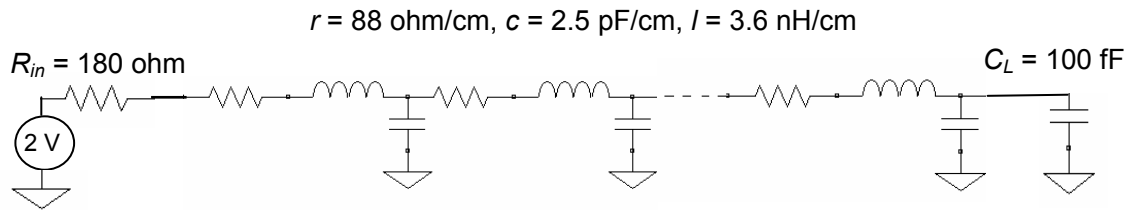


Figure 3.1: Transmission line structure.

3.2.1 RLC bandwidth model

The transmission line model and its equivalent electrical circuit model in [50] are shown in Figure 3.2. Using this electrical circuit model, [50] has given the overall system transfer function as

$$H(j\omega) = \frac{V_o}{V_s} = \frac{(1 - \Gamma_s)(1 + \Gamma_L)}{2} \cdot \frac{H_{\text{inf}}(j\omega)}{1 - \Gamma_s \Gamma_L H_{\text{inf}}(j\omega)^2}, \quad (3.1)$$

where V_o is the output voltage, V_s is the input voltage, and Γ_s and Γ_L are the reflection coefficients at the source and the load, respectively.

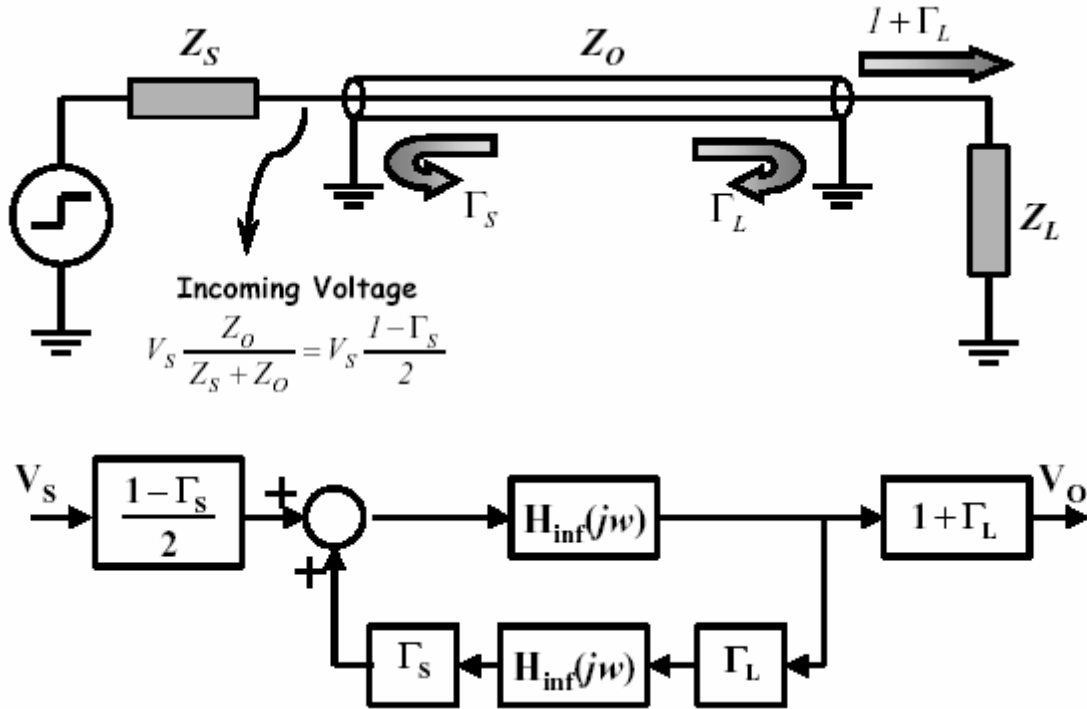


Figure 3.2: Transmission line and its equivalent electrical circuit model in [50].

In [50], $H_{\text{inf}}(j\omega)$, which is the transfer function for an infinite length transmission line, is found to be

$$H_{\text{inf}}(j\omega) = e^{-l z_0 j\omega c}, \quad (3.2)$$

where l is the length of the line, Z_0 is its characteristic impedance, and c is the capacitance per unit length. Substituting (3.2), (3.1) can be simplified as

$$H(j\omega) = \frac{V_0}{V_s} = \frac{(1 - \Gamma_s)(1 + \Gamma_L)}{2} \cdot \frac{e^{-lZ_0j\omega c}}{1 - \Gamma_s\Gamma_L e^{-2lZ_0j\omega c}}, \quad (3.3)$$

which is used to plot the frequency response of the transmission line in [50].

As discussed in [50], in the simplest case, a distributed RC line can be modeled as a low-pass filter (LPF) with a -3 dB bandwidth as

$$BW_{-3dB} = \frac{1}{2\pi RC}. \quad (3.4)$$

It is found in [50] that though the distributed RC model and the distributed RLC model of the transmission line generate different frequency response curves, they have an almost identical -3 dB bandwidth. Therefore, it is assumed that (3.5) can be used to represent the bandwidth of an actual RLC transmission line.

However, the bandwidth expression in (3.5) represents the maximum signal *frequency*, i.e., the inverse of time period, which can be sustained on the transmission line. Throughput T_{max} , which is the inverse of the signal pulsewidth, is therefore twice this bandwidth and is given by

$$T_{max} = \frac{1}{\pi RC}. \quad (3.6)$$

For the transmission line described by parameters in Table 3.1, (3.6) results in a throughput of 1.4 Gbps, which is significantly smaller compared to the actual throughput obtained from HSPICE simulations with small values of R_{in} .

However, it is important to note that HSPICE simulations are performed in this research to achieve a swing of $90\%V_{dd}$ at the output, which translates into -1 dB

attenuation. The -3 dB throughput requires only a $70\%V_{dd}$ swing at the output, and it results in a significantly higher throughput, which further increases the error between the results given by (3.6) and HSPICE simulations. It is seen that the -3 dB throughput model in (3.6) gives significantly pessimistic results compared to HSPICE simulations. However, this behavior is in agreement with the results in [50], which suggests that though (3.4) gives in an identical bandwidth for RC and RLC interconnects, this value is extremely pessimistic compared to the actual values obtained by HSPICE simulations. Nonetheless, (3.6) gives the ultimate lower limit on throughput that can be obtained on a transmission line for a given value of attenuation.

3.2.2 RLC time delay model

The compact time delay models for RLC interconnects are rigorously derived in [51], and their simplified approximate expressions are also presented. The approximate piecewise linear model for the 50% time delay of an RLC interconnect is given in [51] as

$$\begin{aligned}
 \text{Region I: } \frac{R}{Z_0} &\leq 2 \ln \left[\frac{4Z_0}{R_{in} + Z_0} \right] \text{ AND } R_{in} < 3Z_0 \\
 \frac{T_{d,50\%RLC}}{ToF} &= 1.0, \\
 \text{Region II: } \frac{R}{Z_0} &\geq 2 \ln \left[\frac{4Z_0}{R_{in} + Z_0} \right] \text{ OR } R_{in} > 3Z_0 \\
 \frac{T_{d,50\%RLC}}{ToF} &= 0.693 \frac{R_{in}}{Z_0} + 0.377 \frac{R}{Z_0}.
 \end{aligned} \tag{3.7}$$

In (3.7), ToF denotes the time of flight for the transmission line, which is given as

$$ToF = \frac{l}{v}, \tag{3.8}$$

where v is the wave propagation speed in the interconnect dielectric and is calculated as

$$v = \frac{1}{\sqrt{l_{seg}c}}, \quad (3.9)$$

where l_{seg} and c are the inductance and capacitance per unit length, respectively. Using (3.8) and (3.9), ToF is found to be 92 ps, which is in close agreement with the value 84 ps obtained by HSPICE simulations.

Typically, the characteristic impedance Z_0 for the transmission line is given by

$$Z_0 = \sqrt{\frac{R + j\omega L}{j\omega C}}. \quad (3.10)$$

However, for low-loss transmission lines at high frequencies, $R \ll j\omega L$, and Z_0 is simplified to

$$Z_0 \approx \sqrt{\frac{L}{C}}. \quad (3.11)$$

However, (3.11) does not account for the termination impedance. To capture the effect of the load capacitance C_L , (3.11) should be modified to

$$Z_0' \approx \sqrt{\frac{L}{C + C_L}}. \quad (3.12)$$

Using (3.12), (3.7) can be rewritten as

$$\begin{aligned} \text{Region I: } \frac{R}{Z_0'} &\leq 2 \ln \left[\frac{4Z_0'}{R_{in} + Z_0'} \right] \text{ AND } R_{in} < 3Z_0' \\ \frac{T_{d,50\%RLC}}{ToF} &= 1.0, \end{aligned} \quad (3.13)$$

$$\begin{aligned} \text{Region II: } \frac{R}{Z_0'} &\geq 2 \ln \left[\frac{4Z_0'}{R_{in} + Z_0'} \right] \text{ OR } R_{in} > 3Z_0' \\ \frac{T_{d,50\%RLC}}{ToF} &= 0.693 \frac{R_{in}}{Z_0'} + 0.377 \frac{R}{Z_0'}. \end{aligned}$$

Substituting the theoretical value of ToF (92 ps) in (3.13) for Region II, the values of the 50% time delay are shown in Table 3.2. These values are shown for different values of C_L . The simple approximate expressions for the 90% time delay are also presented in [51], which are used to obtain the 90% delay shown in Table 3.2. These values are compared with HSPICE simulation results, which are also shown in Table 3.2. If the expression for Z_0 in (3.11) is used, the resulting values of delay using the models in [51] are shown in Table 3.2. As seen from the highlighted columns in Table 3.2, using Z_0' given by (3.12) gives more accurate results than those obtained using Z_0 given by (3.11), which justifies the inclusion of C_L to the expression of characteristic impedance. It should also be noted that because of the inherent capability of the transmission line to simultaneously carry multiple bits, its throughput for all cases in Table 3.2 is greater than the reciprocal of the 50% or the 90% delay.

Table 3.2: The comparison of delay calculated using models in [51] and HSPICE simulations.

C_L (fF)	50% delay					90% delay				
	HSPICE (ns)	Models in [51] using Z_0	% error	Models in [51] using Z_0'	% error	HSPICE (ns)	Models in [51] using Z_0	% error	Models in [51] using Z_0'	% error
50	0.414	0.382	8.8	0.392	5.4	1.28	1.180	8.1	1.211	5.4
100	0.420	0.382	10.9	0.393	6.5	1.30	1.180	11.2	1.232	5.3
200	0.443	0.382	13.8	0.403	8.9	1.36	1.180	13.3	1.242	8.7
500	0.499	0.382	23.5	0.415	16.6	1.52	1.180	22.4	1.275	16.2

3.2.3 RLC rise time model

The time delay expressions in [51] assume a step input with zero rise time. However, if the input voltage has a finite rise time, the total 90% latency at the output of the transmission line can be approximated as

$$\tau_{90\%} = T_{rise} + T_{d,90\%RLC}, \quad (3.13)$$

where T_{rise} is the 0 to 90% rise time, and $T_{d,90\%RLC}$ is the 90% time delay given in [51] as

$$\begin{aligned} \text{Region I: } \frac{R}{Z_0} &\leq 2 \ln \left[\frac{2.22Z_0}{R_{in} + Z_0} \right] \text{ AND } R_{in} < 1.22Z_0 \\ \frac{T_{d,90\%RLC}}{ToF} &= 1.0, \end{aligned} \quad (3.14)$$

$$\begin{aligned} \text{Region II: } \frac{R}{Z_0} &\geq 2 \ln \left[\frac{2.22Z_0}{R_{in} + Z_0} \right] \text{ OR } R_{in} > 1.22Z_0 \\ \frac{T_{d,90\%RLC}}{ToF} &= 2.3 \frac{R_{in}}{Z_0} + \frac{R}{Z_0}. \end{aligned}$$

To study the effect of input rise time on the delay of the transmission line described in Table 3.1, a ramp input with a finite rise time is used instead of the step input. The actual latency observed from HSPICE simulations and the latency calculated using (3.13) are shown in Figure 3.3 for various values of the input rise time. Figure 3.3 shows that the simple expression in (3.13) compares well with the actual values of latency obtained by HSPICE simulations.

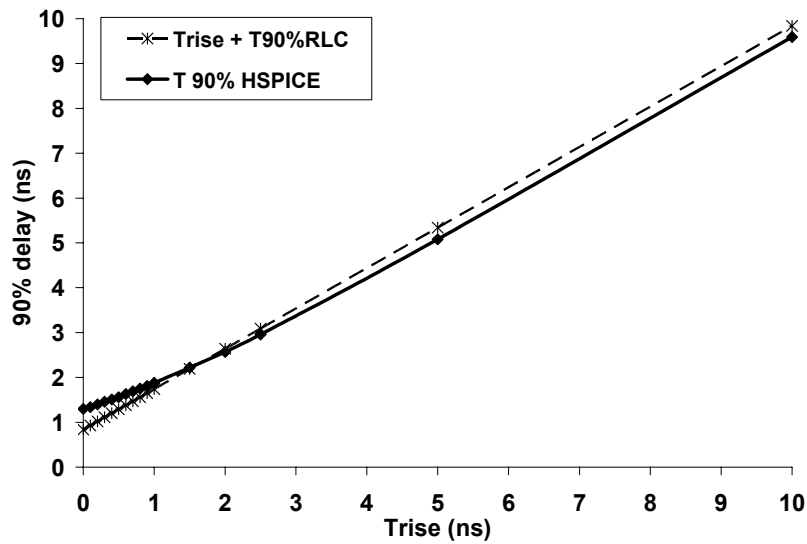


Figure 3.3: Comparison of analytical expression in (3.13) and HSPICE simulation results for transmission line delay with finite input rise time.

It is observed in Figure 3.3 that when the input rise time is significantly larger than the line delay, the total latency is dominated by the rise time. For such cases, similar to [52], the throughput can be approximated as

$$T_{\max} \approx \frac{1}{T_{\text{rise}}}. \quad (3.15)$$

Thus, if the input rise time is significantly greater than the line delay, it can be the primary limiting factor for the maximum bit rate that can be achieved on the line.

3.3 Boundary between RC and RLC models

In different design scenarios, the interconnects can be characterized by different models such as a lumped RC model, a distributed RC model, or a distributed RLC model. A lot of researchers have attempted to define the boundary between the RC and RLC characterization regimes for the interconnect circuits [35], [53], [54]. Naeemi has provided some interesting insights into the modeling of coplanar RLC interconnects [35]. Based on the delay-centric approach for an interconnect with a square cross-section, [35] presents the values of the interconnect width that define the boundary between the RC and RLC regime. It is shown in [35] that the characterization of an interconnect as an RC interconnect results in infinitesimally small error in the interconnect delay when the interconnect width is less than this boundary width. Therefore, [35] advocates that accounting for the inductance becomes more and more important as the interconnect dimensions become larger.

With technology scaling, though the transistor parameters are scaled down, the global interconnects are reverse-scaled to mitigate the interconnect bottleneck problem.

Therefore, in general, the consideration of inductance may become necessary to accurately characterize the behavior of global interconnects in future technology generations.

However, the characterization of repeater-inserted interconnects could be significantly different from that of a low-loss VLSI transmission line because the insertion of repeaters plays an important role in the characterization of the interconnect circuits. Venkatesan et al. have presented interesting insights into this problem in [46]. The expressions for the latency of RC and RLC interconnects with repeaters are presented in [46], which lead to the unified expression for the latency as

$$t_{d,rep} = \max \left(ToF, 0.377 \frac{rcl^2}{n} + 0.693 \frac{R_0 cl}{h} \right) + 0.693 C_0 h \left(rl + 0.65n \frac{R_0}{h} + 0.36nZ_0' \right), \quad (3.16)$$

where r and c are the interconnect resistance and capacitance, respectively, R_0 and C_0 are the resistance and capacitance of a minimum-sized driver, respectively, and n and h are the number and size of repeaters, respectively. Based on (3.16), [46] suggests that when the time of flight for an interconnect segment is greater than the RC charging time for its distributed capacitance, the interconnect behavior is inductive. Otherwise, it is resistive. Simplifying (3.16) further, this result is expressed in [46] as

$$0.377 \frac{rl}{nZ_0'} + 0.693 \frac{R_0}{hZ_0'} \begin{cases} \leq 1: \text{Inductive, RLC regime.} \\ > 1: \text{Resistive, RC regime.} \end{cases} \quad (3.17)$$

The expression of Z_0' in (3.17) can be written as

$$Z_0' \approx \sqrt{\frac{L_{seg}}{C_{seg} + C_t}}, \quad (3.18)$$

which is modified from (3.12) for a repeater-inserted interconnects. It is seen from (3.17) that with the insertion of repeaters, the value of the numerator rapidly decreases, but because of the presence of C_t , the value of the denominator decreases slowly. As a result, Z_0' decreases as more repeaters are inserted on the interconnect. The values of Z_0' for different repeater densities for the interconnect described in Table 3.1 are shown in Table 3.3.

Table 3.3: Values of Z_0' for different repeater densities.

n	1	5	10	20
Z_0' (ohm)	36	34	31	27

Because of a larger associated coefficient, the second term in (3.17) is more dominating. Consequently,

$$\text{for } \frac{R_0}{hZ_0'} > 1.443, \quad 0.377 \frac{rl}{nZ_0'} + 0.693 \frac{R_0}{hZ_0'} > 1. \quad (3.19)$$

For the 180 nm technology node with a 2 V supply voltage, R_0 is found to be approximately 10 Kohm. The value of Z_0' is found to be 36 ohm for a single-driver interconnect, which further decreases with the insertion of repeaters as seen in Table 3.3. Even with Z_0' of 36 ohm, h needs to be less than 192 for (3.19) to hold true and make the RC characterization of the interconnect satisfactory. The values of h that are greater than 192 are seldom used in any practical applications. Therefore, even if the first term in (3.17) reduces with the insertion of repeaters, the second term alone is significantly larger than unity, which holds the wave-pipelined interconnect in the RC regime. Moreover, the insertion of repeaters makes the effective segment resistance (i.e., segment resistance + transistor resistance) more dominant compared to the characteristic impedance, which dampens the impact of inductance on interconnect performance.

It is seen using (3.17) that the first term, which is an order of magnitude smaller than the second term for ten repeaters, becomes two orders of magnitude smaller than the second term for 20 repeaters. Thus, the dominance of the second term in (3.17) significantly increases with the insertion of more repeaters. Because Z_0' reduces with the insertion of repeaters, it increases the value of the already dominant second term in (3.17), which further pushes the interconnect into the RC regime. Therefore, (3.17) suggests that the repeater-inserted interconnects can be modeled as RC interconnects with a minimal error in the results for most practical applications.

The accuracy of the RC characterization of the wave-pipelined interconnect is also validated using HSPICE simulations. The HSPICE results for throughput and latency for the RLC interconnect are compared to those for the RC interconnect in Table 3.4, for different repeater densities. It is seen in Table 3.4 that the percent errors in the values of throughput and latency using RC models w.r.t. the RLC models decrease as the number of repeaters increases, which supports the inferences drawn earlier in this section.

Table 3.4: Comparison between performance of RC and RLC wave-pipelined interconnect circuits using HSPICE simulations.

Number of repeaters	Throughput			Latency			Value of metric and regime according to (3.17)
	RLC (Gbps)	RC (Gbps)	% error	RLC (ns)	RC (ns)	% error	
1	1.18	1.09	7.31	0.451	0.454	0.67	4.38, RC
5	3.03	2.99	1.49	0.651	0.653	0.31	3.88, RC
10	4.55	4.51	0.88	0.806	0.808	0.25	4.13, RC
20	5.95	5.90	0.84	1.101	1.102	0.09	4.68, RC

3.4 Comparison of bit rates obtained on transmission line and wave-pipelined interconnect

Having discussed the impact of inductance on the interconnect performance and interconnect characterization, this section presents the comparison between bit rates obtained on the transmission line and the wave-pipelined interconnect circuit. In theory, the maximum bit rate on the lossless transmission line can be infinite. However, it is the finite resistance of the transmission line and the input resistance that limit the maximum bit rate that can be achieved on it. Though the inherent capacity of the transmission line for high-speed serialization increases its throughput beyond the reciprocal latency, this enhancement is also primarily limited by the input resistance.

Table 3.5 shows the HSPICE results for throughput and 90% latency for the transmission line described in Table 3.1, which is terminated using a 100 fF capacitance. This capacitance is equivalent to the gate capacitance of a transistor with a scaling factor of 56, which is used for most of the analysis in this research. To see the impact of the input resistance on the line performance, R_{in} is varied from 0 ohm to 180 ohm. (The equivalent resistance of a transistor having a scaling factor of 56 is 180 ohm.) The values of the reciprocal latency are also shown in Table 3.5 to highlight the fact that a new bit is transmitted on the line before the previous bit reaches a 90% voltage swing because an inherent data parallelism exists on the line.

Table 3.5: Impact of input resistance on transmission line performance.

R_{in}	Throughput	90% Latency	90% reciprocal latency
0	7.7 Gbps	0.19 ns	5.26 Gbps
5	6.8 Gbps	0.23 ns	4.35 Gbps
50	2.2 Gbps	0.50 ns	2.00 Gbps
120	1.2 Gbps	0.93 ns	1.08 Gbps
180	0.8 Gbps	1.31 ns	0.76 Gbps

It is seen in Table 3.5 that even if a large driver with an equivalent resistance of 50 ohm is used, a 2.2 Gbps throughput is obtained on the transmission line. An equal or higher throughput can be easily obtained on the wave-pipelined interconnect, which has same dimensions as the transmission line, by inserting more repeaters. HSPICE simulations show that a 2.2 Gbps throughput can be obtained by wave-pipelining the interconnect with only three repeaters with $h = 56$. The 90% latency of the wave-pipelined interconnect is then 0.68 ns, which is not significantly degraded compared to some of its values obtained on the transmission line in Table 3.5. Moreover, to achieve an equivalent resistance of 50 ohm, the transmission line needs to be driven by a repeater that is more than three times as big as that used on the wave-pipelined interconnect. As a result, the wave-pipelined interconnect would occupy comparable area and dissipate almost the same power as the transmission line, without any loss of throughput performance.

The saturation throughput of the wave-pipelined interconnect, calculated using (2.30), is 6.8 Gbps. To obtain this throughput on the transmission line, an input resistance of 5 ohm is needed. The driver size of 2000 is needed to achieve an equivalent resistance of 5 ohm, which is practically impossible to obtain. Therefore, wave-pipelining can achieve a significantly higher throughput than that can be obtained on the transmission line in any practical applications.

The 180 nm transmission line is assumed to have a wire pitch of 2 μm , which is 15% larger than that used for the Intel 180 nm global interconnects [29]. Therefore, the transmission line results in a significantly large wire area. However, a 2.2 Gbps throughput that is obtained on the transmission line with an input resistance of 50 ohm

can also be achieved on the wave-pipelined interconnect having smaller dimensions. For instance, HSPICE simulations show that the same throughput can be obtained on an interconnect with a 1 μm wire pitch by inserting 12 repeaters with $h = 56$. *The wave-pipelined interconnect thus achieves the required throughput at the expense of half the wire area.* This example clearly shows that a throughput identical to or more than that obtained on the transmission line can be easily obtained on the wave-pipelined interconnect at the expense of less power and/or area. The power and area for the three design choices discussed in this section for a 2.2 Gbps throughput are summarized in Table 3.6. It is seen that wave-pipelining either reduces power or area compared to the transmission line, for the same throughput performance.

Table 3.6: Transmission line and wave-pipelining design choices for a constant throughput performance.

Design technique	Transmission line	Wave-pipelining	
		Interconnect dimensions unchanged	Interconnect dimensions changed
Throughput	2.2 Gbps		
h	200*	56	56
Silicon area	3.14E-06 cm^2	2.64E-06 cm^2	10.56E-06 cm^2
Wire area	2E-04 cm^2	2E-04 cm^2	1E-04 cm^2
Dynamic power	13.04 mW	12.72 mW	17.86 mW

* $h = 200$ for $R_{in} = 50 \text{ ohm}$

Thus, it is seen in this section that the throughput of the transmission line is limited by the *input resistance* and it cannot be enhanced beyond this limitation for a given value of input resistance. However, this bit rate can be increased on the wave-pipelined interconnect by adding more repeaters. High-speed data serialization and periodic boosting of data signals achieved by wave-pipelining enhances the interconnect throughput beyond that can be obtained on practical transmission lines. Moreover, as

seen in the examples in this section, wave-pipelining renders the designer a large design space that offers good design flexibility. It also facilitates the use of several variations of the repeater insertion technique. Though uniform buffer insertion is primarily covered in this research, the use of inverting repeaters or cascaded inverter repeaters are some other options that can be explored to enhance throughput or reduce power or area. *Thus, wave-pipelining can not only outperform the transmission line, but it also gives the designer a better control over choosing the interconnect and transistor dimensions to meet various design constraints.*

3.5 Impact of wave-pipelining on RLC crosstalk

3.5.1 Simulation of RLC interconnect circuits using HSPICE and RAPHAEL

The interconnect inductance can not only affect the throughput and latency but it can also affect the crosstalk. The inductive and capacitive coupling between adjacent interconnects can cause noise transients on quiet lines, voltage overshoots on active lines, or variations in the throughput and latency. In this section, all these effects are collectively referred to as ‘crosstalk’.

To analyze the impact of wave-pipelining on the RLC crosstalk, considering a single interconnect is not sufficient. Parallel wire channels need to be considered for the crosstalk analysis, and all the mutual inductances and capacitances need to be accurately modeled. Therefore, an interconnect system that contains five interconnects and two non-ideal ground planes is considered, and all the mutual capacitances and inductances are accurately calculated using RAPHAEL RC-2 models.

The capacitance parasitics for a 5 interconnect, 2 ground plane system are shown in Figure 3.4. The lines at two ends are assumed to be ground lines. The self and the mutual capacitances for the central interconnect are denoted by C_g and C_m , respectively. As seen in Figure 3.4, C_g is calculated w.r.t. the ground planes above and below the interconnect. In a real system, these ground planes could be replaced by orthogonal data lines, and the impact of orthogonal lines on the wave propagation speed is discussed in [35]. However, assuming the existence of ground planes is a good first-order approximation for calculating the interconnect capacitance [35], which is used in this research.

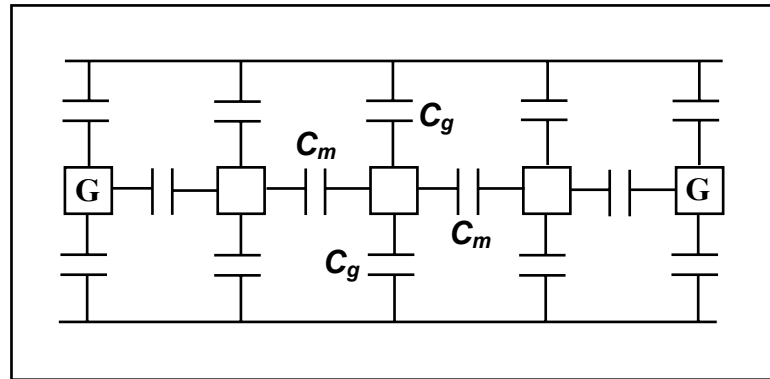


Figure 3.4: A 5 interconnect, 2 ground plane system with self and mutual capacitances.

Similar to [35], to make the calculation of inductance more realistic, the ground planes are removed from the interconnect system. For calculating the interconnect inductance, it is not necessary to assume the presence of orthogonal lines because the current does not return through these lines. The return paths for the current exist in the ground lines in the vicinity of the data line. The exact current distribution among several ground lines on the same tier depends on the interconnect geometry and the frequency of

operation. However, for the interconnects carrying high-frequency currents that are considered in this research, the return path is assumed to be in the ground lines at two ends. The interconnect system used to calculate the inductance using RAPHAEL is shown in Figure 3.5.

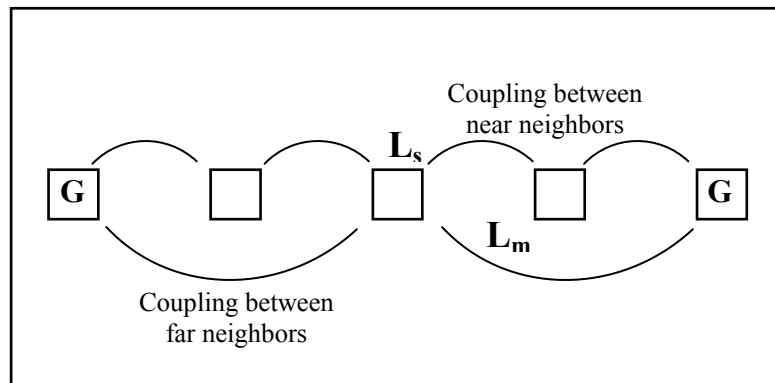


Figure 3.5: A 5 interconnect system with self and mutual components of inductance.

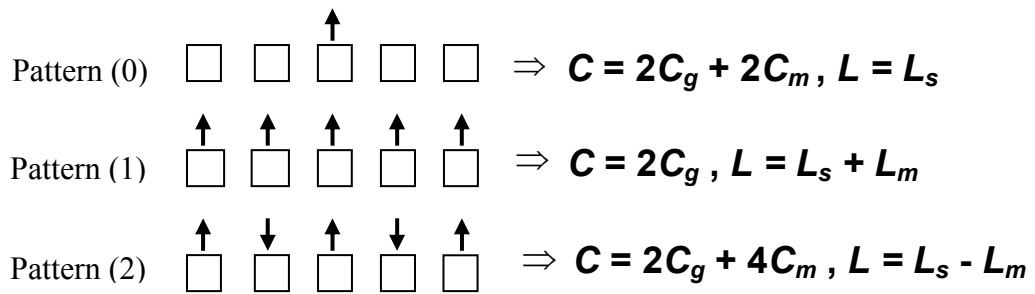
The self and mutual capacitances and inductances calculated using RAPHAEL are used to model the distributed RLC interconnect system in HSPICE. One RLC segment is used for 0.01 cm interconnect length. HSPICE simulations using level-49 transistor models [48] are then performed for an interconnect with and without repeaters to analyze the impact of wave-pipelining on the interconnect crosstalk.

3.5.2 Impact of wave-pipelining on dynamic delay effects, overshoot voltage, and crosstalk voltage in RLC interconnect systems

Because parallel wire channels are linked to each other through the mutual components of the interconnect parasitics, the switching patterns have a significant impact on the throughput, latency, and crosstalk. This subsection briefly discusses the

inductive and capacitive crosstalk and dynamic delay effects resulting from different switching patterns.

When the adjacent interconnects exhibit different switching patterns, two nodes of a mutual capacitance undergo different voltage swings. Therefore, the contribution of the mutual capacitance to the total interconnect capacitance is different for different switching patterns. Similarly, depending on the switching patterns, the electromagnetic fields resulting from the currents in the adjacent interconnects are either additive or subtractive, which changes the contribution of the mutual inductance to the total inductance. Figure 3.6 shows three different switching patterns for an interconnect system with five signal lines. In pattern (0), only the middle interconnect switches and the neighbors are quiet. In pattern (1), all the interconnects simultaneously switch in the same direction, whereas in pattern (2), the adjacent interconnects switch simultaneously in the opposite direction. The equivalent interconnect capacitance and inductance for all three patterns are also shown in Figure 3.6.



C : Total capacitance, C_g : Ground capacitance, C_m : Mutual capacitance
 L : Total inductance, L_s : Self inductance, L_m : Mutual inductance

Figure 3.6: Equivalent capacitances and inductances for different switching patterns for a 5 interconnect, 2 ground plane system.

It is seen in Figure 3.6 that the equivalent capacitance of the interconnect is maximum when its neighbors simultaneously switch in the opposite direction, i.e., pattern (2). From (2.21), it can be concluded that pattern (2) would result in the lowest throughput among these three patterns. However, the mutual inductance shows the exact opposite behavior w.r.t. the switching patterns than the mutual capacitance. The electromagnetic fields resulting from the currents are additive for pattern (1) and they are in the opposite directions for pattern (2). Therefore, if L_s is the self inductance of the interconnect and L_m is the mutual inductance, the equivalent inductance for pattern (1) is given by L_s+L_m and that for pattern (2) is given by L_s-L_m .

HSPICE simulations are performed to analyze the impact of the inductive and capacitive coupling on the throughput, latency and crosstalk. The interconnect length is 1 cm and other dimensions are same as those in Table 3.1. However, the interconnect and repeater parameters are rewritten in Table 3.7 for convenience.

Table 3.7: Parameters used in HSPICE simulations for crosstalk analysis.

Supply voltage	2 V	Repeater size	56
Ground capacitance	0.25 pF	Mutual capacitance	1 pF
Resistance	88 ohm	Self inductance	3.6 nH
Inductance coupling factor – near neighbors	0.5738	Inductance coupling factor – far neighbors	0.3317

Restricting the overshoot voltage on the active interconnects to small values is important for maintaining good signal integrity and reliability. A large overshoot spike on the active line can cause more crosstalk voltage to appear on a neighboring quiet line or worsen the reliability of the chip by damaging the gate oxide [35]. Therefore, smaller values for the overshoot voltage are desirable on the active interconnects. Table 3.81

shows the values of throughput, latency, and active line overshoot voltage for the parameters shown in Table 3.7, for the switching patterns (1) and (2).

Table 3.8: HSPICE Results for throughput, latency, and active line overshoot voltage.

Switching Pattern	(1) Simultaneous switching in the same direction			(2) Simultaneous switching in the opposite direction		
Number of Repeaters	Throughput (bps)	Latency (ns)	Overshoot Voltage	Throughput (bps)	Latency (ns)	Overshoot Voltage
1	5.55E+09	0.15	160 mV	6.25E+08	0.52	100 mV
2	6.15E+09	0.21	200 mV	1.11E+09	0.49	500 mV
5	6.18E+09	0.41	200 mV	1.67E+09	0.65	400 mV
8	6.20E+09	0.59	150 mV	2.00E+09	0.83	300 mV
10	6.23E+09	0.72	80 mV	2.27E+09	0.96	300 mV
15	6.25E+09	1.02	80 mV	2.63E+09	1.27	220 mV
20	6.27E+09	1.31	80 mV	2.86E+09	1.56	200 mV
25	6.29E+09	1.61	80 mV	2.94E+09	1.85	200 mV
30	6.29E+09	1.89	80 mV	3.08E+09	2.15	180 mV

The interconnect parameters shown in Table 3.7 indicate that severe inductive and capacitive coupling exists for the interconnect geometry chosen for this analysis. As seen earlier in this section, the inductive crosstalk is more dominant in the event of simultaneous switching in the same direction, and the capacitive crosstalk is more dominant when the adjacent interconnects switch simultaneously in the opposite direction. The results in Table 3.8 indicate that the throughput is limited more by capacitive coupling. As a result, the throughput is significantly less for switching pattern (2) than pattern (1).

The overshoot voltage, on the other hand, is a result of inductive coupling. Table 3.8 shows a trend in which the active line overshoot voltage initially increases with the insertion of a few repeaters because of the faster data transitions. However, as the repeater density further increases, the impact of inductance is dampened further, which

reduces the active line overshoot voltage (by a maximum of 50%), thereby improving the signal integrity.

An observation of Table 3.8 shows that the dynamic delay effects resulting from the inductive and capacitive coupling are more severe for the low-loss transmission line with a driver (i.e., the single-driver interconnect) than the wave-pipelined interconnect. For the low-loss transmission line, there is a difference of an order of magnitude between the values of throughput for the two switching patterns. This difference is considerably lower for the same interconnect with 30 repeaters. Similarly, the relative percent differences in the latency are significantly higher for the low-loss transmission line than the wave-pipelined interconnect. This fact highlights the importance of wave-pipelining to reduce the dynamic delay effects.

Figure 3.7 shows the impact of wave-pipelining using repeaters on the quiet line crosstalk voltage. To obtain these results, the middle interconnect is assumed to be quiet (Q) and the neighbors simultaneously switch in the same ($\uparrow\uparrow$) or opposite ($\uparrow\downarrow$) direction. The fact that the inductive coupling is more dominant in the case of crosstalk is highlighted by the results in Figure 3.7 because simultaneous switching of the neighbors in the same direction results in a significantly higher crosstalk voltage. In this case, the insertion of repeaters reduces the crosstalk by 75% compared to the low-loss transmission line. Figure 3.7 shows that wave-pipelining using repeaters helps maintain good signal integrity by reducing the crosstalk voltage on quiet lines. Thus, wave-pipelining can be effectively used to reduce both the active line overshoot voltage and the quiet line crosstalk voltage.

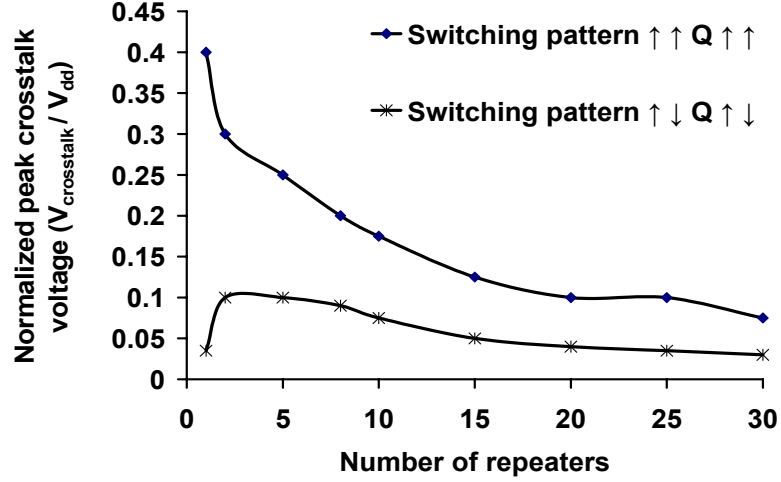


Figure 3.7: Normalized crosstalk voltage on a quiet line.

3.6 Impact of wave-pipelining on power supply noise

To analyze the impact of wave-pipelining on power supply noise (PSN), the interconnect dimensions identical to the previous section are considered. The interconnect circuit is assumed to be driven by a 2 V supply with $\pm 10\%$ variations. The maximum frequency component in the PSN is assumed to be twice the nominal frequency of operation of that interconnect circuit. A single low-loss transmission line is compared to the interconnect having identical dimensions, but wave-pipelined using 10 inverting repeaters, to study the impact of PSN on throughput and latency. A driver scaling factor of 56 is used. The nominal values of the throughput and latency with an ideal power supply are shown in Table 3.9.

Table 3.9: Nominal values of throughput and latency for low-loss transmission line and wave-pipelined interconnect circuit.

Type of interconnect	Low-loss transmission line	Wave-pipelined interconnect
Number of repeaters per cm	1	10
Nominal throughput	1.163 Gbps	7.143 Gbps
Nominal 50% latency	0.375 ns	0.971 ns

The histograms in Figure 3.8 and Figure 3.9 show the results for throughput and latency over 1000 HSPICE simulations for the transmission line and the wave-pipelined interconnect, respectively. Based on these histograms for random PSN, the values of the absolute errors in the throughput and latency for the low-loss transmission line and the wave-pipelined interconnect circuit are shown in Table 3.10. These errors are a result of the power supply variations and are calculated w.r.t. the nominal values of the throughput and latency in Table 3.9. It is seen from Table 3.10 that wave-pipelining significantly reduces the absolute errors compared to the low-loss transmission line, thereby ensuring a stable performance. Table 3.9 shows that wave-pipelining using 10 repeaters results in more than six times increase in the throughput performance compared to the low-loss transmission line, and Table 3.10 shows that this high performance is achieved with a relatively smaller error.

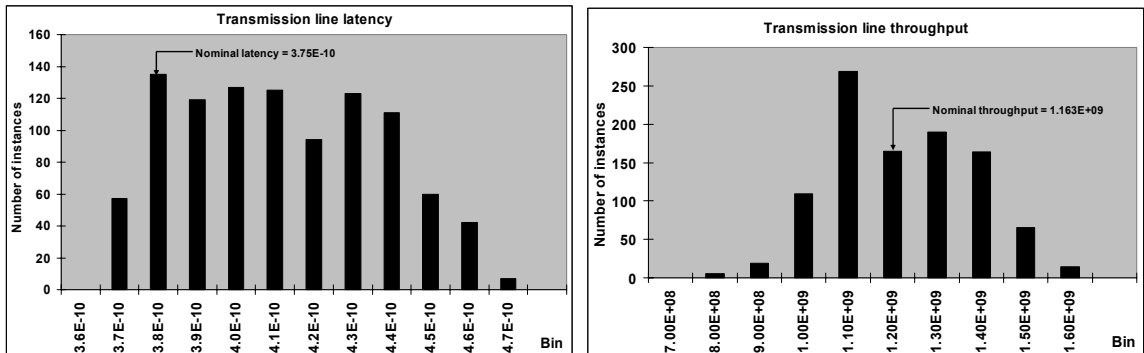


Figure 3.8: Histograms for throughput and latency of transmission line.

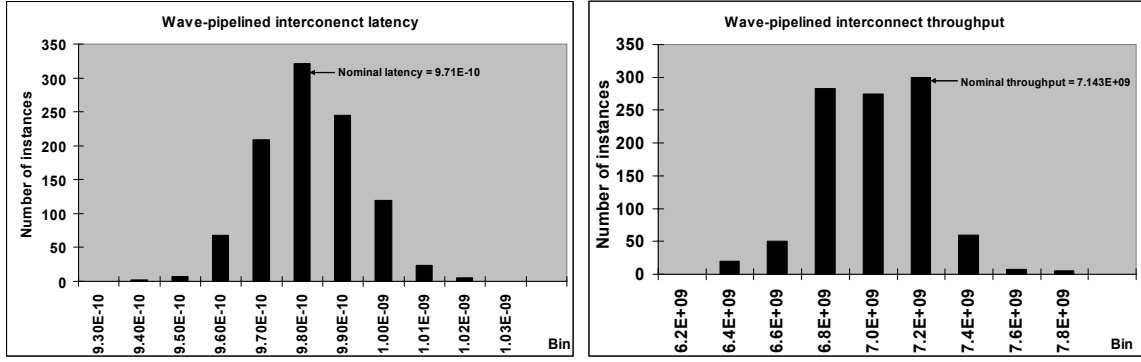


Figure 3.9: Histograms for throughput and latency of wave-pipelined interconnect.

Table 3.10: Absolute errors for low-loss transmission line and wave-pipelined interconnect.

Type of error	Low-loss transmission line		Wave-pipelined interconnect	
	Throughput	Latency	Throughput	Latency
Average absolute error	9.68%	8.69%	3.02%	1.21%
Maximum absolute error	34.35%	24.32%	12.50%	4.87%

Even though the low-loss transmission line results in a smaller value of the latency than the wave-pipelined interconnect, it results in larger latency variations. On the other hand, the maximum error in the latency of the wave-pipelined interconnect circuit, over 1000 simulations, is only $\sim 5\%$, which can be tolerated by the output latches.

On the wave-pipelined interconnect, the power supply fluctuations seen by one repeater driver get balanced with those seen by the other. Moreover, the periodic signal boosting provided by repeater circuits helps filter glitches and retain a stable performance. On the other hand, for the low-loss transmission line, because there are no intermediate repeater drivers, a power supply fluctuation seen at any instant ripples through the line to the output, thereby causing larger performance variations. Therefore, to summarize, wave-pipelining not only enhances the performance and the signal integrity but it also shows better tolerance to PSN by reducing performance variations.

3.7 Techniques to minimize performance variations on wave-pipelined interconnect circuits

As seen in Table 3.8, the throughput performance greatly depends on the switching patterns in an interconnect network. Therefore, the maximum interconnect throughput is limited to the lower of the two throughput values corresponding to the two switching patterns. The switching patterns also cause variations in the communication latency, which could cause serious problems in receiving and synchronizing the data on wave-pipelined interconnects. As seen in Section 3.6.2, though wave-pipelining dampens the impact of inductive coupling, dynamic delay effects resulting from capacitive coupling are significant. Therefore, techniques to minimize the delay and throughput variations should be employed for wave-pipelined interconnect circuits.

3.7.1 Shielding ground lines insertion

Inserting more shielding ground lines is one solution to minimize the performance variations. If the signal lines are surrounded by ground lines, the ground lines provide the necessary return paths to the high-frequency currents. The ground lines can provide the necessary shielding from the excessive inductive and capacitive coupling and dampen the impact of switching patterns on the interconnect performance and signal integrity.

For instance, if every signal line is surrounded by one ground line on each side, the mutual capacitance between the signal line and the neighboring ground lines always remains constant regardless of the switching patterns on other signal lines. Because the total capacitance of the line remains unchanged, (2.21) shows that the throughput remains constant and becomes independent of the switching patterns on other signal lines.

These ground lines also act as a shield from inductive coupling. It is seen in Table 3.7 that the inductive coupling factors between near neighbors are significantly higher than those between far neighbors. The insertion of a shielding ground line between two signal lines makes the signal lines far neighbors of each other, thereby reducing the inductive coupling between them. As a result, there is a reduction in the crosstalk, which results in better signal integrity.

Moreover, if one ground line is inserted per signal line, the total wire area is identical to that of a twisted pair of differential interconnects. However, because the ground lines significantly reduce the impact of inductive and capacitive coupling on the interconnect performance, inserting ground lines gives an opportunity to reduce wire pitch and still achieve the necessary performance at the expense of smaller wire area.

3.7.2 Misaligned repeater insertion

Misaligned repeater insertion proposed in [5], [14] is another useful technique to minimize performance variations. As shown in Figure 3.10, the staggered insertion of *inverting* repeaters results in identical switching with the neighbor over half of the interconnect length and opposite switching over the other half of the interconnect length, which makes the performance less dependent on the switching patterns.

As shown in Table 3.11, HSPICE simulations show that misaligned repeater insertion can reduce the throughput variation from 82% to 25% and latency variation from 60% to 9% for a 180 nm global interconnect with dimensions similar to those in earlier sections ($w = s = 1 \mu\text{m}$, $h = t = 2.5 \mu\text{m}$). For the interconnects that do not exhibit as much inductive and capacitive coupling, i.e., a 180 nm metal-5 interconnect in [26],

misaligned repeater insertion is found to result is less than 10% variation in both throughput and latency. With misaligned repeater insertion, the chosen global interconnect with a 2 μm pitch can be designed to achieve a throughput of 1.667 Gbps instead of 1.428 Gbps, as seen in Table 3.11.

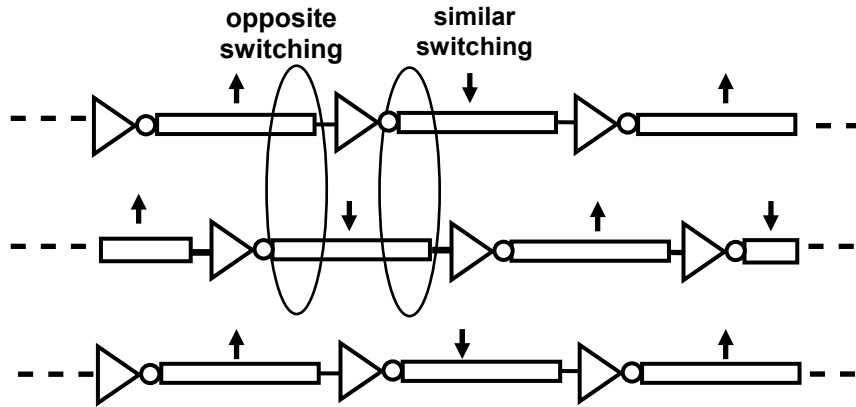


Figure 3.10: Interconnects with misaligned repeaters [5].

Table 3.11: HSPICE results for a 180 nm global interconnect with four repeaters.

Technique	Uniform repeater insertion		Misaligned repeater insertion	
Switching pattern	Throughput (bps)	Latency (s)	Throughput (bps)	Latency (s)
Pattern (0) <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> </div>	2.273E+09	5.71E-10	2.273E+09	5.71E-10
Pattern (1) <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> </div>	4.167E+09	3.23E-10	2.857E+09	5.19E-10
Pattern (2) <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↓ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↓ </div> <div style="border: 1px solid black; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center;"> ↑ </div> </div>	1.428E+09	9.20E-10	1.667E+09	6.14E-10

It is shown in [14] that the optimal positions for misaligned repeaters could be different from the midpoint of the adjacent segments. The expressions for these optimal positions are calculated in [14] for minimum delay and noise on bidirectional buses. It is shown in [14] that the optimum repeater positioning provides a significantly lower noise pulse amplitude and lower sensitivity of propagation delay and noise pulse peak to segment length variation, compared to commonly used midway repeater positioning.

The concept of misaligned repeater insertion for the repeater-inserted interconnects is similar to staggered twisting of the differential interconnects [55]. However, staggered twisting requires a careful routing and placement analysis of the interconnect networks, whereas misaligned repeater insertion is relatively easier to apply to the on-chip interconnect circuits.

3.7.3 Decoupling capacitor insertion

Inserting decoupling capacitors is a commonly used technique to reduce the impact of simultaneous switching noise (SSN) on circuit performance and signal integrity [56]-[58]. Decoupling capacitors are typically inserted in the white spaces available on the chip. The decoupling capacitors, which are inserted between the two nodes of a power supply, minimize the impact of power supply noise on signal integrity by temporarily storing the power supply voltage on them and returning it to the circuit when needed. Depending on their value, decoupling capacitors also allow certain frequency components to pass through them to the ground, thereby filtering the noise at those frequencies.

A detailed analysis of the allocation and placement of decoupling capacitors at the floorplan level is performed in [56]. The insertion of decoupling capacitors is shown to result in up to a 75% reduction in the peak noise for certain benchmarks in [56]. Because of the area-centric approach, the insertion of decoupling capacitors in [56] also results in significantly larger amounts of unused white spaces on the chip. To calculate the values of the decoupling capacitance for a particular application, a combination of time-domain analysis and frequency-domain analysis is suggested in [57]. The capacitance required

near the clock frequency is determined by time-domain calculations, whereas the capacitance required at other frequencies is determined by circuit simulation in the frequency domain in [57].

3.8 Summary

The impact of inductance on interconnect performance is analyzed in this chapter through a discussion of various performance models for RLC interconnects. The existing performance models are analyzed to capture the transmission line behavior. The boundary between the RC and the RLC regimes is discussed in this chapter to obtain insights into modeling of low-loss VLSI transmission lines and repeater-inserted interconnects. It is shown that wave-pipelined interconnects can be accurately characterized as RC interconnects for their performance analysis because the resistance of repeaters significantly dampens the impact of inductance, thereby pushing the wave-pipelined interconnect circuits into the RC regime.

It is shown that wave-pipelining using repeaters can not only enhance the interconnect throughput beyond that obtained on the low-loss transmission line, but it also offers the designer a large design space, which increases the design flexibility and creates an opportunity to reduce area and power. Wave-pipelining is also shown to be a more effective technique to maintain good signal integrity and reduce performance variations in the presence of severe inductive and capacitive coupling and power supply fluctuations, compared to the transmission line. To summarize, wave-pipelining dampens the impact of inductance on interconnect performance and significantly enhances the performance and signal integrity compared to the low-loss transmission line.

CHAPTER 4

VOLTAGE SCALING REPEATER INSERTION (VSRI) CIRCUIT ANALYSIS

4.1 Introduction

This chapter discusses the importance of supply voltage scaling to reduce the total power dissipation. The simultaneous application of voltage scaling and repeater insertion (VSRI) is proposed in this chapter for low-power, high-performance interconnects. The impact of VSRI on the throughput performance and signal integrity is analyzed in this chapter. VSRI is also compared to low-voltage differential signaling (LVDS), which is another well-known technique for low-power, high-performance interconnects, in the fields of power and area for an identical throughput performance. Moreover, the impact of power supply fluctuations, which is one of the major sources of on-chip noise, on both VSRI and LVDS is analyzed with the help of HSPICE simulations. Finally, the redesign of the VSRI circuit is presented in this chapter to mitigate the impact of power supply fluctuations on the interconnect throughput and latency.

4.2 Importance of voltage scaling

As a result of a large number of transistors and a high operating frequency, ITRS projects significantly large values of power dissipation for the upcoming technology generations [59]. Increased transistor count along with an increase in design complexity also results in a proportional increase in the number of interconnects on the chip. Therefore, the power dissipated by interconnect circuits can be a significant portion of the total power dissipation on the chip [17], [60]. Because of the increased parasitic capacitance of repeaters, wave-pipelining could further increase the power dissipation on the interconnect circuits.

The primary contribution to the total power comes from the dynamic power for technology nodes up to 100 nm [12]. The dynamic power can be modeled as

$$P_d = \alpha \left(\frac{1}{2} C_{total} V_{swing}^2 f \right), \quad (4.1)$$

where α is the switching activity of the circuit, C_{total} is the total switching capacitance, V_{swing} is the voltage swing that the circuit undergoes (which is equal to the supply voltage in most cases), and f is the frequency of operation. To reduce the dynamic power, one or more of α , C_{total} , V_{swing} , and f need to be reduced.

The terms α and C_{total} are dependent upon the technology generation and the circuit parameters, and controlling these terms could require significant changes at the circuit and the architecture level. However, a few researchers have proposed low-power techniques based on the reduction of α [20], [21]. The term f represents the frequency of operation, and reducing f reduces the performance; however, this is not desirable in any high-performance applications. However, a careful observation of (4.1) shows that the

dynamic power varies with the square of the voltage swing, and a significant power reduction can be achieved by scaling down this voltage swing that the circuit undergoes.

Low-voltage differential signaling (LVDS) is one signaling standard used in the industry that was developed to obtain high speed on low-power interconnects. In this technique, a reduced voltage swing of a few hundreds of milli-volts is used on the differential interconnect [18], [61]. The interconnect undergoes a smaller voltage swing, which helps reduce its power and enhance its speed. The differential signaling used in LVDS results in the elimination of common mode noise, but the use of a smaller voltage swing could make LVDS interconnects more susceptible to single-ended noise and other dynamic delay effects [62]. A large wire area, susceptibility to single-ended noise, and increased design complexity are some drawbacks of this technique.

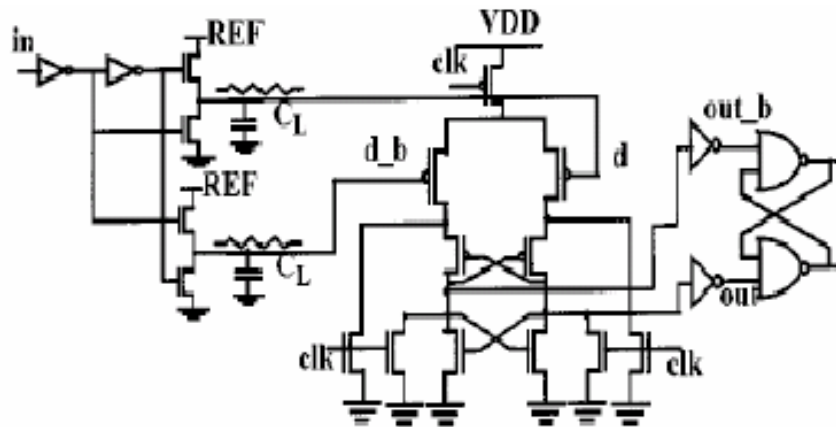


Figure 4.1: Circuit implementation of LVDS [18].

Figure 4.1 shows the actual circuit implementation of LVDS [18]. A slight variation of this circuit is also shown in [63], [64]. As mentioned in Section 1.1.2, some other low-voltage-swing techniques such as pulse-controlled drivers [18] and symmetric driver and level converters for low-power ULSIs [19] can also be found in the literature.

4.3 Simultaneous application of voltage scaling and repeater insertion (VSRI)

Even though LVDS and other low-swing techniques mentioned in Section 1.1.2 use a low voltage swing on the interconnect to reduce power, they do not use a scaled supply voltage. These low-swing techniques use complex drivers, receivers, and level converters, which operate at a large supply voltage and dissipate a lot of power. The total power dissipation in these complex driver-receiver circuits overshadows the interconnect power reduction for shorter interconnect lengths.

The leakage power and the short-circuit power, which are predicted to make a significant contribution to the total power for the future technology generations [12], vary directly with the supply voltage. Therefore, using a scaled supply voltage for the entire interconnect circuit reduces the leakage and the short-circuit power of the driver-receivers, in addition to the total switching power. Moreover, the use of a scaled supply voltage for the entire interconnect circuit also makes its design simpler. Therefore, supply voltage scaling is critical to reduce the total power dissipation.

Despite its advantages, rigorous supply voltage scaling is not practiced because it results in a lower drive current that in turn deteriorates circuit performance. However, the interconnects can be wave-pipelined using repeaters to effectively recuperate the performance loss resulting from voltage scaling. The simultaneous application of (supply) voltage scaling and repeater insertion (VSRI) can thus reduce the power dissipation without any loss of throughput performance and is a very useful technique for high-performance, low-power interconnect networks.

For wave-pipelined *logic* circuits, supply voltage scaling results in a larger delay for each pipelined stage, which deteriorates the circuit performance. Because of the

rigidity of the logic design, it is difficult to split one stage into more stages to reduce the stage delay and enhance the throughput. Larger transistors may be used to enhance performance, but they increase the load on the previous stage and may give only a marginal increase in performance. On the other hand, for wave-pipelined *interconnect* circuits, one stage can be easily split into more stages by inserting more repeaters to enhance the throughput. Because of relatively relaxed design constraints, voltage scaling can be effectively used on wave-pipelined interconnect circuits to reduce power, without any loss of throughput performance.

Figure 4.2 shows the variation of throughput with the number of repeaters, with the supply voltage scaled from 2 V to twice the threshold voltage. It is seen in Figure 4.2 that the throughput drops with supply voltage scaling, as expected. However, for a given value of the supply voltage, the throughput is significantly enhanced by repeater insertion before it saturates to a constant value. The operation in the saturation region translates into a large number of repeaters without a significant increase in throughput. Therefore, to gain the maximum benefits from this technique, it is important to operate near the knee of the throughput curve (before the saturation region), which from the inspection of Figure 4.2 is seen to be 10-20 repeaters per unit cm.

It is seen in Figure 4.2 that the same throughput can be obtained by using various combinations of supply voltages and numbers of repeaters. Therefore, some optimization analysis is necessary to decide the ideal combination for a given throughput. For instance, it is seen in Figure 4.2 that a 2.1 Gbps throughput can be achieved by using a combination of 1.5 V and 10 repeaters/cm (design A) or by another combination of 1 V and 20 repeaters/cm (design B). Though design B results in a 100% increase in the silicon

area because of its doubled repeater count compared to design A, it reduces the power by almost 40% because of the lower supply voltage. Understanding this area-power trade-off helps optimize VSRI circuits to achieve high performance and low power.

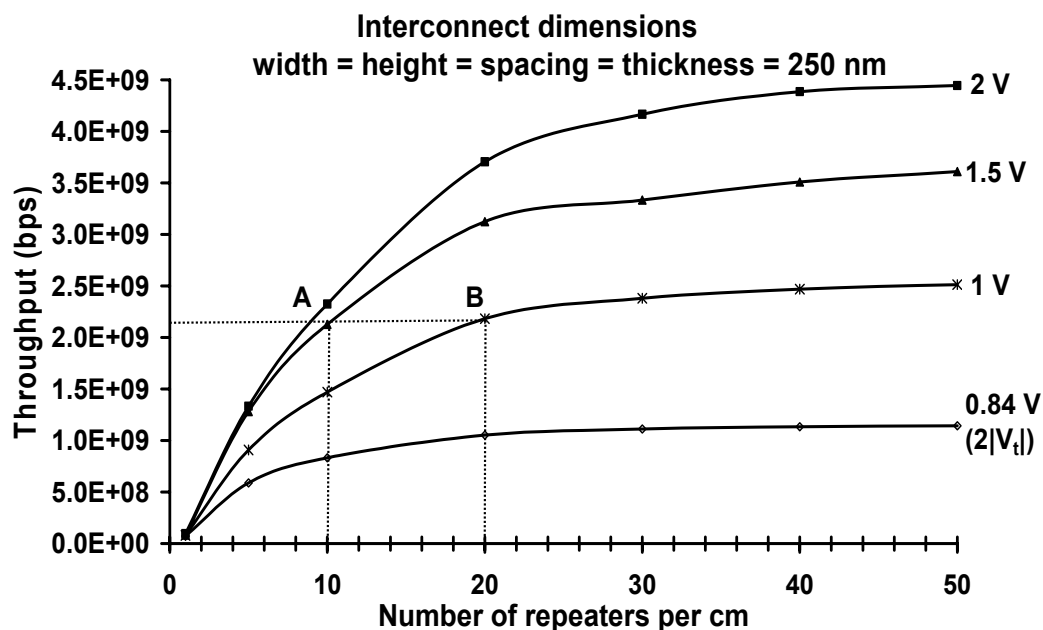


Figure 4.2: Effect of VSRI on throughput.

As discussed earlier in this section, scaling the supply voltage reduces the current drive to the circuit and makes the circuit slower, thereby increasing the latency. The insertion of repeaters beyond the optimal design point in [6] also increases the latency. Therefore, VSRI, in general, tends to increase the latency, as seen in Figure 4.3. However, because the interconnects are wave-pipelined, an increase in latency does not translate into a loss of throughput performance. It is seen in Figure 4.3 that a VSRI design using a 1 V supply and 10 repeaters results in the same latency as a single-driver interconnect operating with a 2 V supply, but it increases the throughput by more than 20 times.

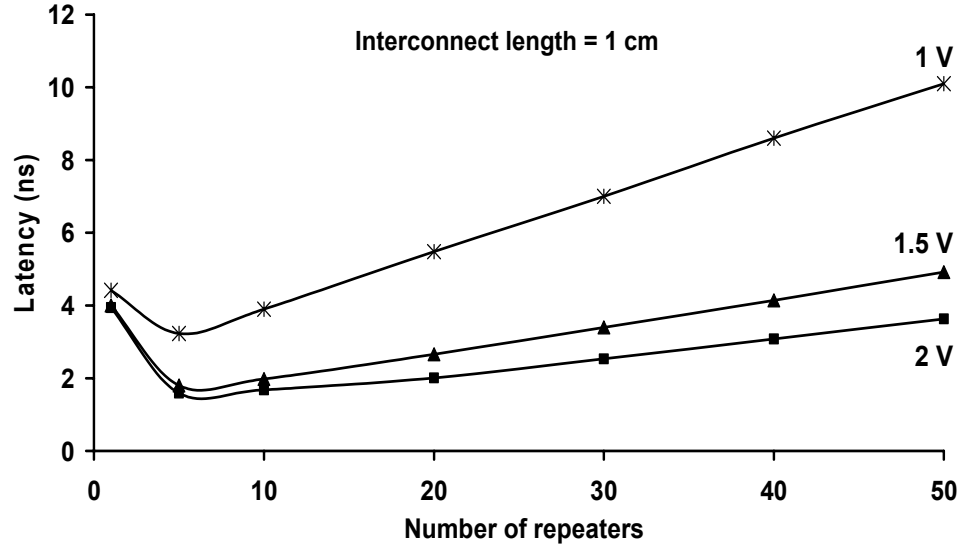


Figure 4.3: Effect of VSRI on communication latency.

For the applications where the latency of the first data bit is very important, a combination of the optimal latency-centric design suggested in [6] and wave-pipelining can be used. The latency-sensitive interconnects can be designed to operate with an optimal number and size of repeaters to minimize the communication latency [6], but instead of limiting throughput to the reciprocal of latency, interconnects can be wave-pipelined to achieve a higher throughput. This idea is further developed in Section 5.8.1.

4.4 Impact of VSRI on signal integrity

The impact of VSRI on the interconnect performance is studied in the previous section, and this section analyzes the impact of VSRI on signal integrity. Similar to Chapter 3, the signal integrity analysis is performed by studying the active line overshoot voltage and the quiet line crosstalk voltage. It is observed in Section 3.6.2 that signal integrity is affected more by the inductive coupling. Therefore, a 5-interconnect system in

which all the interconnects switch simultaneously in the same direction is considered to model the dominant inductive coupling.

4.4.1 Impact of VSRI on overshoot voltage

Table 4.1 shows HSPICE results for the active line overshoot voltage normalized to the supply voltage for a 1 cm long interconnect having a 2 μm pitch, as the supply voltage is scaled from 2 V to 1 V. The interconnect parameters are same as those shown in Table 3.7. A repeater scaling factor of 56 is used. It is seen from the results in Table 4.1 that voltage scaling only marginally changes the values of the active line overshoot voltage. The trend shown by Table 4.1 suggests that using a lower supply voltage and a larger repeater density slightly reduces the overshoot voltage compared to that at a high supply voltage and fewer repeaters. Thus, Table 4.1 shows that VSRI certainly does not worsen the overshoot voltage compared to the conventional interconnect design, but it slightly improves it.

Table 4.1: Active line overshoot voltage normalized to supply voltage ($V_{overshoot}/V_{dd}$) for different supply voltages and repeater densities.

Supply voltage	Number of repeaters per cm				
	1	2	5	10	20
2 V	0.075	0.100	0.100	0.040	0.040
1.5 V	0.060	0.067	0.070	0.060	0.060
1 V	0.080	0.060	0.070	0.065	0.065

4.4.2 Impact of VSRI on crosstalk voltage

Keeping all the transistor and interconnect parameters unchanged, Table 4.2 shows the results of HSPICE simulations for the quiet line crosstalk voltage normalized

to the supply voltage, as the supply voltage is scaled from 2 V to 1 V. The neighbors of the quiet line are assumed to switch simultaneously in the same direction to simulate the maximum inductive coupling scenario. It is seen in Table 4.2 that VSRI significantly reduces the crosstalk voltage on the quiet line. A single-driver interconnect operating with a 2 V supply results in the maximum crosstalk voltage, whereas an interconnect operating with a 1 V supply and 20 repeaters results in the minimum crosstalk voltage in the chosen design range.

Voltage scaling lowers the magnitude of the currents and voltages in the interconnect circuit, thereby lowering the crosstalk. Repeater insertion dampens the impact of inductive coupling on interconnect as seen in Section 3.6.2, which further reduces crosstalk. Therefore, VSRI is a very effective technique to maintain good signal integrity on the interconnects.

Table 4.2: Quiet line crosstalk voltage normalized to supply voltage ($V_{crosstalk}/V_{dd}$) for different supply voltages and repeater densities.

Supply voltage	Number of repeaters per cm				
	1	2	5	10	20
2 V	0.400	0.300	0.250	0.175	0.109
1.5 V	0.310	0.295	0.200	0.141	0.105
1 V	0.250	0.240	0.180	0.138	0.100

A throughput of ~ 2 Gbps can be achieved by using a 2 V supply and 9 repeaters per cm, a 1.5 V supply and 10 repeaters per cm, or a 1 V supply and 19 repeaters per cm. The table entries roughly corresponding to these design points are highlighted in Table 4.2. It is seen in Table 4.2 that in addition to reducing the power dissipation, the design using a lower supply voltage and a larger number of repeaters also reduces crosstalk,

without any loss of throughput performance. This fact underlines the importance of VSRI to maintain good signal integrity.

4.5 Comparison of VSRI and LVDS

Both VSRI and LVDS are designed to obtain high performance and low power on global interconnect networks. The motivation behind using LVDS is its rejection of the common mode noise, whereas for VSRI, it is its ability to achieve significantly high performance through high-speed serialization using wave-pipelining. Therefore, VSRI and LVDS are compared in this section in the areas of performance, power, area, and noise immunity.

4.5.1 Circuit configurations for LVDS and VSRI

The configuration of the LVDS circuit is shown in Table 4.3 based on [63]. The configuration of the VSRI circuit that uses the same wire pitch and achieves the same throughput as LVDS is also shown in Table 4.3. The transistor technology corresponds to the 180 nm node. The interconnect length is 0.5 cm because [63] optimizes and validates the LVDS circuit for this interconnect length.

Table 4.3: LVDS and VSRI circuit configurations for a 0.5 cm long 180 nm interconnect.

Design technique	LVDS [63]	VSRI
Interconnect dimensions	$w = 0.8 \mu\text{m}$, $s = 0.8 \mu\text{m}$, $h = 1.6 \mu\text{m}$, $t = 1.6 \mu\text{m}$	$w = 0.4 \mu\text{m}$, $s = 1.2 \mu\text{m}$, $h = 1.0 \mu\text{m}$, $t = 1.0 \mu\text{m}$
Design configuration	1.6 V supply, 600 mV differential swing	1 V supply, 2 repeaters / 0.5 cm
Transistor scaling factor	5 – 55	56

4.5.2 LVDS and VSRI comparison for performance, power, and area

HSPICE simulations using level-49 MOSIS transistor models in [48] show that both the configurations shown in Table 4.3 achieve a 2 Gbps throughput on the 0.5 cm shielded interconnect. The performance, power, and area of these circuits are shown in Table 4.4. It is seen in Table 4.4 that for the same performance, VSRI results in more than a 40% reduction in the total power, a 50% reduction in the wire area, and a 15% reduction in the latency.

Table 4.4: Comparison between LVDS and VSRI for a 2 Gbps throughput.

Design technique	LVDS	VSRI
Throughput	2 Gbps	
Total power	3.466 mW (Dynamic: 3.22 mW, Leakage: 0.35 μ W, Short-circuit: 0.25 mW)	1.293 mW (Dynamic: 1.241 mW, Leakage: 0.55 μ W, Short-circuit: 0.05 mW)
Wire area for data lines	16E-05 cm ²	8E-05 cm ²
Silicon area	8.9E-07 cm ²	17.6E-07 cm ²
Latency	0.783 ns	0.664 ns

The total power shown in Table 4.4 consists of the switching power, leakage power (subthreshold and gate), and short-circuit power. The use of a lower supply voltage and single-ended signaling results in considerably lower power for VSRI compared to LVDS. Moreover, though the LVDS and VSRI circuits use an identical wire pitch, differential signaling in LVDS requires double wire area than VSRI, which makes VSRI more suitable for wire-limited applications compared to LVDS. VSRI also results in a lower latency than LVDS, which is important for latency-sensitive applications.

Even though the VSRI circuit requires more silicon area for the 0.5 cm interconnect analyzed in this section, the silicon area scales linearly for larger interconnect lengths and assures the same throughput. For example, for a 1 cm long interconnect, doubling the number of repeaters would double the silicon area of the VSRI circuit, but it would assure a 2 Gbps throughput. If LVDS needs to be used on a 1 cm long interconnect, doubling the transistor sizes does not guarantee that the same throughput can be obtained. The LVDS circuit needs to be redesigned and optimized every time the interconnect length changes. Consequently, there could be an interconnect length where LVDS either does not even achieve the required throughput or requires much larger silicon area than VSRI to achieve the required throughput.

4.5.3 Impact of power supply noise on LVDS and VSRI

Power supply noise, which is also referred to as the simultaneous switching noise (SSN), is one of the greatest sources of noise for on-chip interconnects [56], [58], [65]. Because of the simultaneous switching activities on the chip, the power supply voltage fluctuates around its regulated DC value. The fluctuations in the supply voltage affect the normal operation of the circuit and can cause significant performance variations [66]. The variation in the latency can cause problems in synchronizing the data at the output of the interconnects. For wave-pipelining, the variation in the throughput could result in possible intersymbol interference (ISI), thereby making the interconnect system unstable. Therefore, the sensitivity to power supply noise is one important design criterion for on-chip interconnect circuits.

To analyze the impact of power supply noise on VSRI and LVDS, HSPICE simulations are performed, where a power supply variation of $\pm 10\%$ is randomly generated around V_{dd} . The ground is assumed to be a perfect zero [66]. The highest frequency component in the noise waveform is twice the operational frequency [66], i.e., 4 GHz in this experiment. Using HSPICE, 1000 simulations are performed for each of LVDS and VSRI to ensure that all the frequency components up to 4 GHz appear in the noise waveforms with an almost equal probability. An HSPICE timing diagram for power supply noise for a 1 V supply is shown in Figure 4.4.

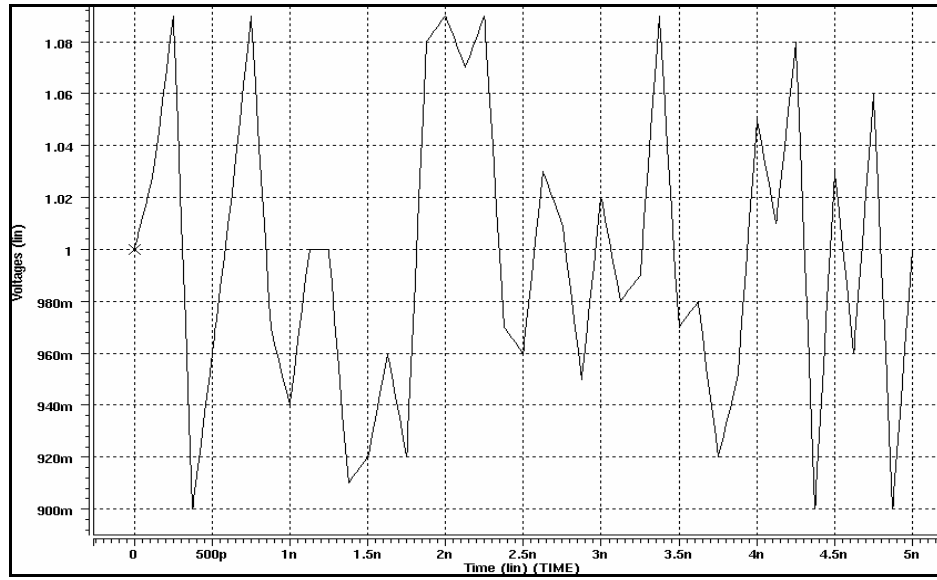


Figure 4.4: Example of supply noise with maximum frequency component of 4 GHz.

The nominal values of the throughput and latency with an ideal power supply are shown in Table 4.4. The absolute errors in the throughput and latency are observed with respect to (w.r.t.) these nominal values. Table 4.5 shows the values for the average and

maximum absolute error for both LVDS and VSRI. The absolute error corresponding to a certain instance of power supply noise is defined as

$$\text{Absolute error} = \frac{|\text{actual value} - \text{nominal value}|}{\text{nominal value}}. \quad (4.2)$$

The histograms corresponding to the values of the throughput and latency for both LVDS and VSRI are presented in Figure 4.5 and Figure 4.6, respectively. In these histograms, the frequency corresponding to a particular bin is the number of instances between the present value and the previous value of the bin label.

Table 4.5: Average and maximum values of absolute error for LVDS and VSRI.

Quantity	Throughput		Latency	
Type of circuit	LVDS	VSRI	LVDS	VSRI
Average absolute error	5.73%	6.61%	2.76%	5.29%
Maximum absolute error	24.24%	26.47%	10.71%	25.50%

It is seen from Table 4.5 and Figure 4.5 that both LVDS throughput and VSRI throughput are almost equally susceptible to power supply noise. LVDS shows slightly better results than VSRI because of the rejection of common mode noise by differential signaling. The values of maximum absolute error in Table 4.5 show that supply voltage fluctuations would require that both LVDS and VSRI operate at almost 25% lower throughput than the nominal throughput. It is seen in Figure 4.5 that for both LVDS and VSRI, almost 70% cases result in lower values of throughput than its nominal value of 2 Gbps. Based on 1000 simulations, this result translates into a 70% probability of failure for a throughput requirement of 2 Gbps.

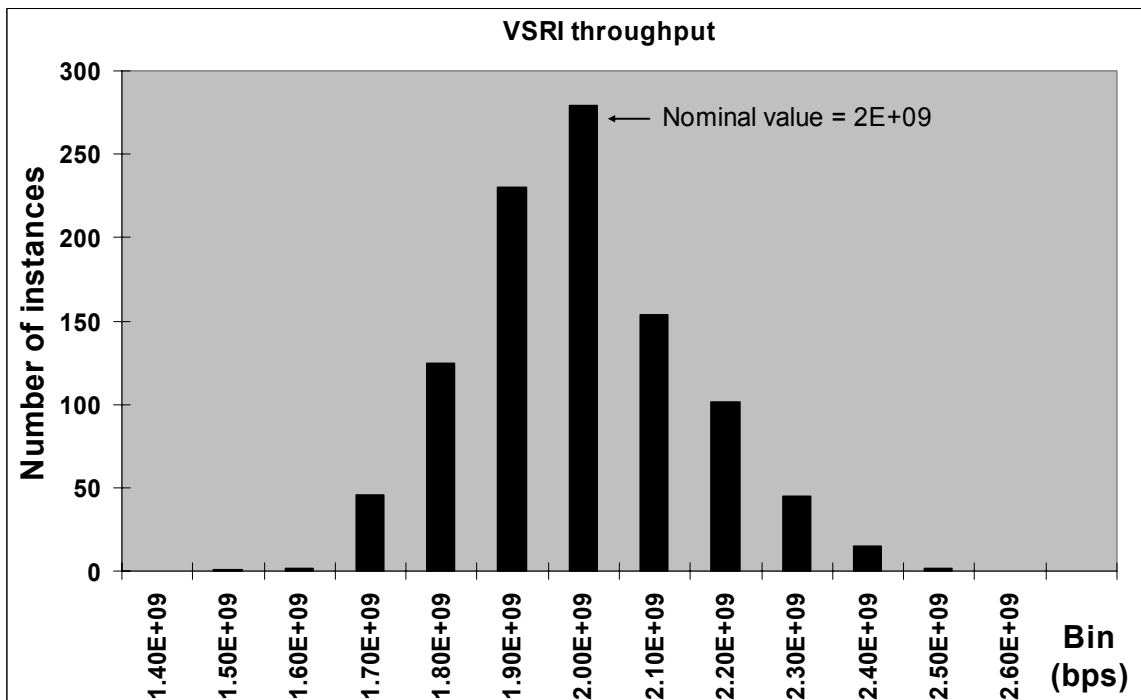
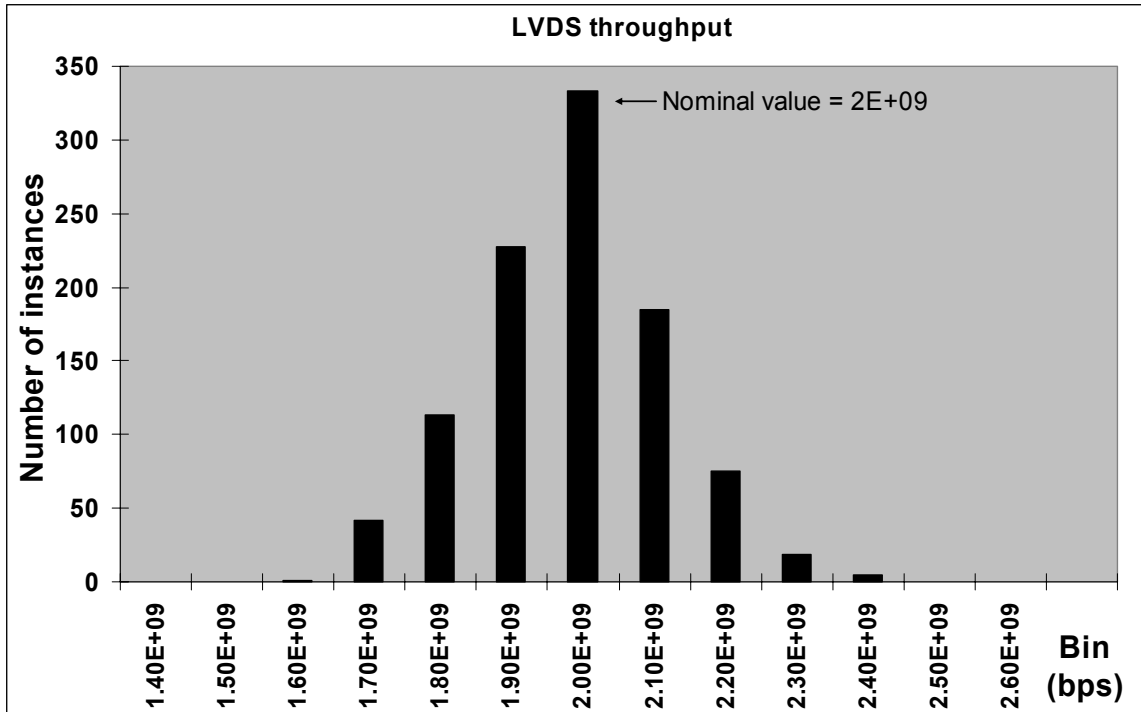


Figure 4.5: Histograms for values of throughput for LVDS and VSRI, over 1000 simulations.

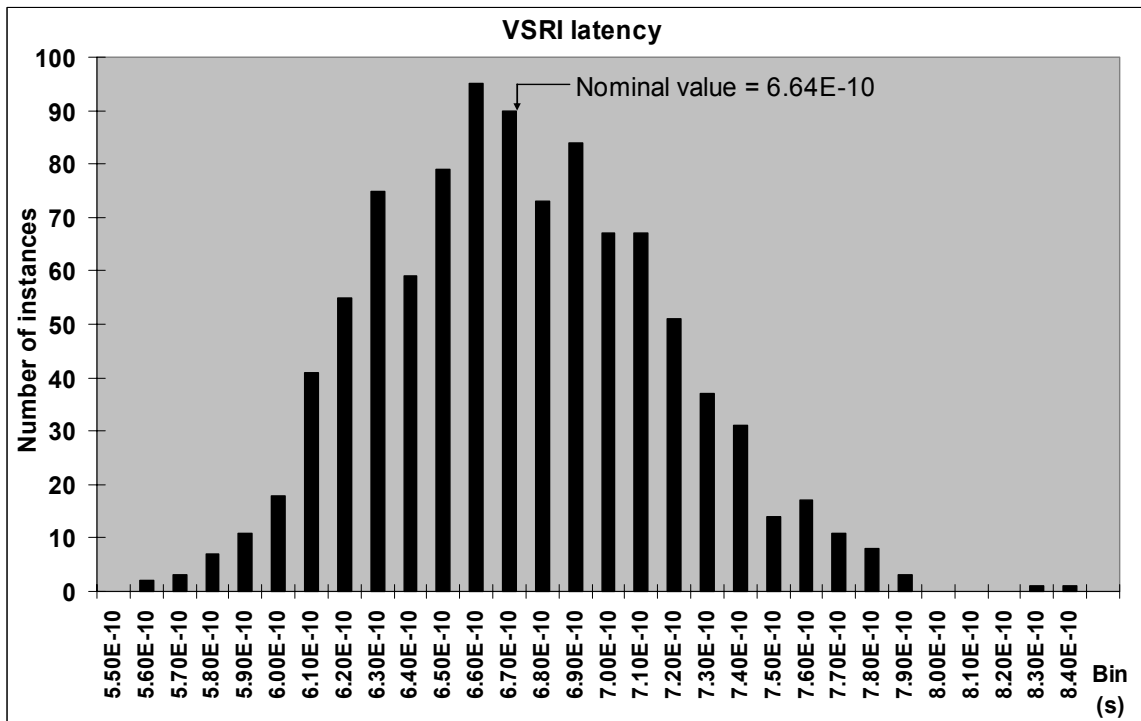
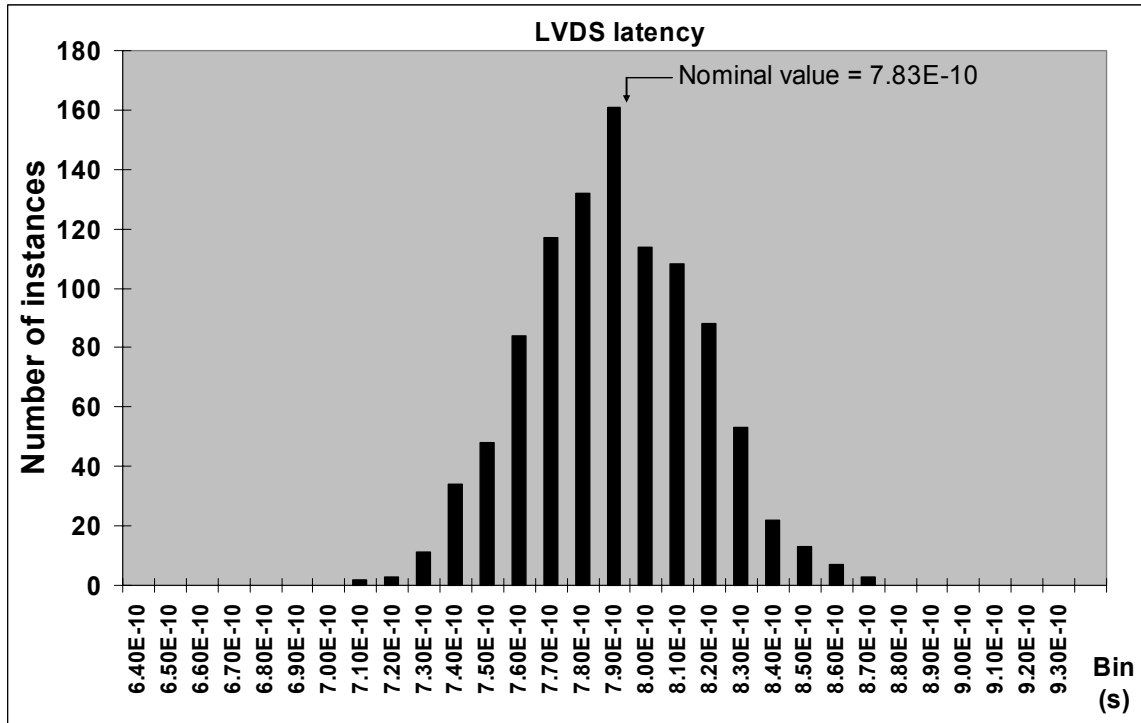


Figure 4.6: Histograms for values of latency for LVDS and VSRI, over 1000 simulations.

However, in case of latency, the variation on either side of the nominal value could be equally detrimental as it could disrupt the operation of the timing circuit. It is seen in Figure 4.6 that LVDS results in relatively less variations in the latency compared to VSRI. Table 4.5 shows that the maximum error in the latency is much smaller in the case of LVDS than VSRI. Therefore, it may seem relatively easier to synchronize the data to clock at the output of LVDS. However, the clock can be sent along with data in the case of VSRI, which would make the data synchronization in VSRI circuits almost insensitive to the latency variations. This synchronization scheme is further discussed in Chapter 6.

It is important to note that unlike LVDS, the throughput performance of the wave-pipelined interconnect circuits using VSRI is independent of the latency. However, for LVDS, the variation in the latency not only affects the operation of the synchronization circuit, but it also affects the interconnect performance. Therefore, any variation in the latency can be more detrimental for LVDS than VSRI.

4.5.4 VSRI overdesign for stable performance

It is seen earlier in Figure 4.2 that the throughput increases with an increase in the repeater density. Making use of this fact, if more repeaters are inserted on the interconnect, the maximum throughput that can be obtained on the VSRI interconnect increases. When the interconnect circuit is capable of operating at a significantly higher throughput than the required throughput, the required throughput can be certainly achieved even in the presence of power supply noise.

To further explain this idea, it is assumed that a 2 Gbps throughput is required on the 0.5 cm interconnect. However, to achieve this throughput, instead of two, five repeaters are inserted on the interconnect. With an ideal supply voltage, this design can give a maximum nominal throughput of 3 Gbps and a latency of 1.185 ns. The histograms for the throughput and latency in the presence of power supply noise for the VSRI design point of 5 repeaters per 0.5 cm interconnect length are shown in Figure 4.7, based on 1000 HSPICE simulations. The absolute errors in the throughput and latency are shown in Table 4.6.

The minimum throughput obtained by this design point over 1000 HSPICE simulations is 2.273 Gbps, which is significantly higher than the required throughput of 2 Gbps. None of the 1000 random instances of power supply noise reduces the VSRI throughput below 2 Gbps. Therefore, it can be safely assumed that if this design configuration is used to obtain a 2 Gbps throughput, it can almost always achieve it despite the presence of power supply noise. Thus, *overdesign is the key to avoid performance variations*. Table 4.7 shows the comparison between this new VSRI design with the existing LVDS design for a 2 Gbps throughput. It is seen in Table 4.7 that the price to pay for this overdesign strategy is that the latency is almost doubled compared to the original VSRI design; however, if throughput is the desired goal, this could be a reasonable tradeoff for certain kinds of interconnects.

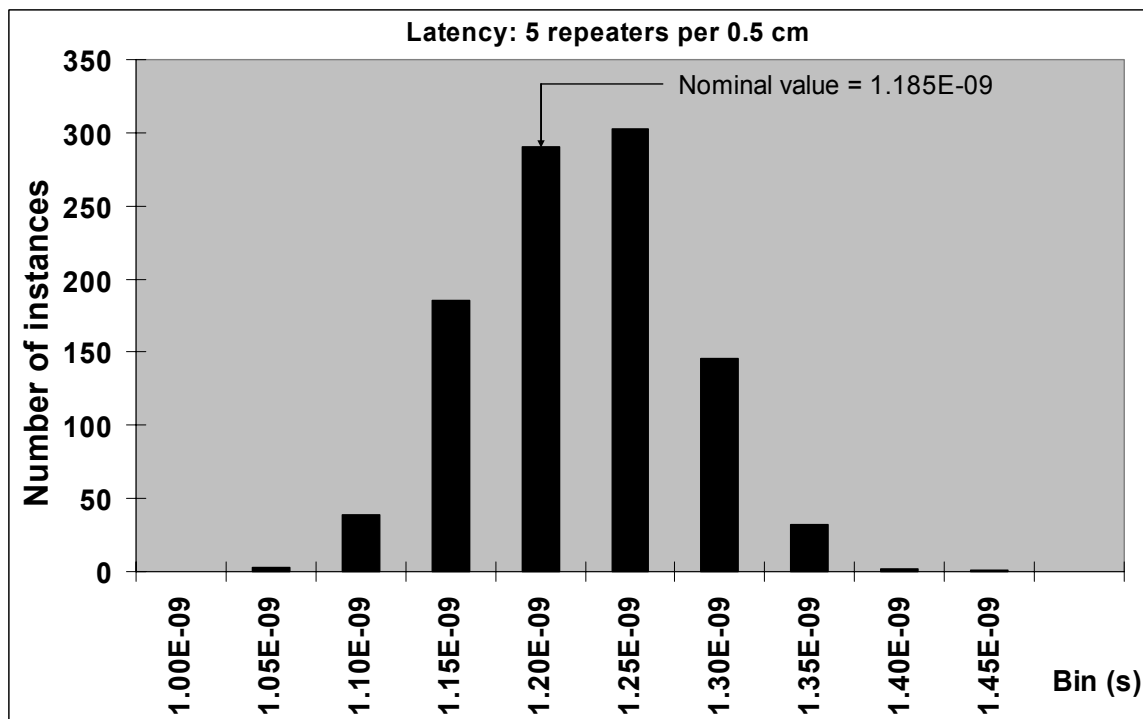
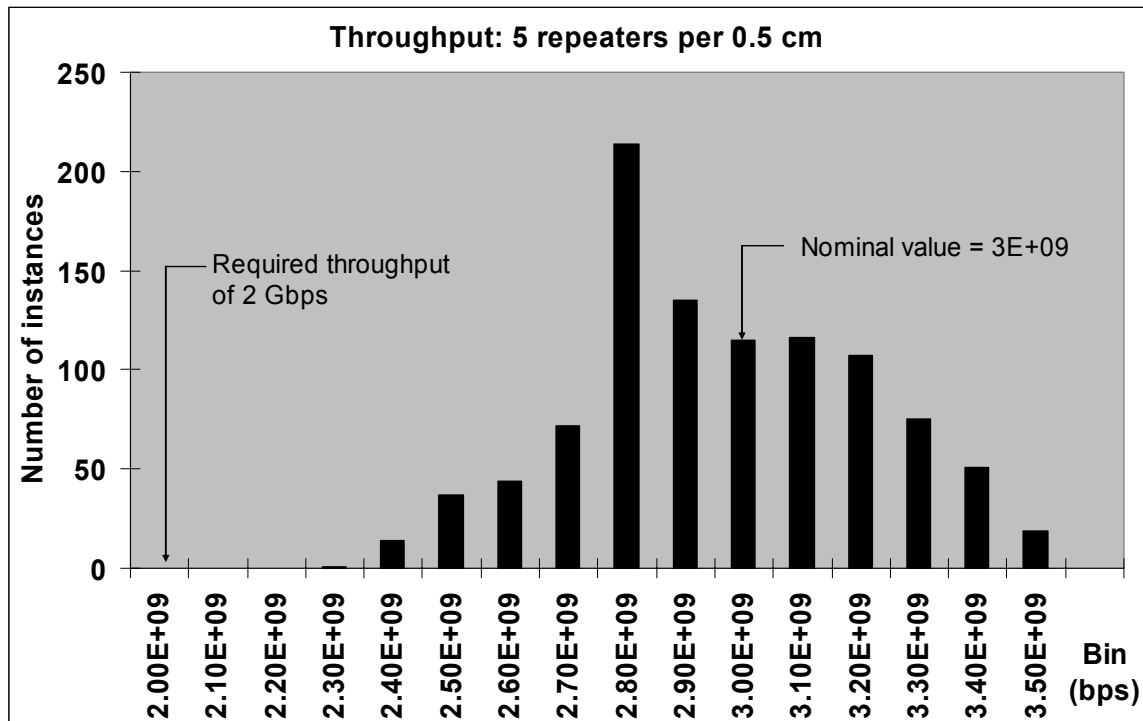


Figure 4.7: Histograms for throughput and latency for a VSRI design point of 5 repeaters per 0.5 cm interconnect length.

Table 4.6: Absolute errors for a VSRI design point of 5 repeaters per 0.5 cm.

Type of error	Throughput	Latency
Average absolute error	7.15%	4.01%
Maximum absolute error	24.32%	18.57%

Table 4.7: Comparison between new VSRI design and existing LVDS design.

Design type	LVDS [63]	VSRI
Design configuration	1.6 V supply, 600 mV interconnect swing, 1.6 μm pitch, differential signaling	1 V supply, 5 repeaters per 0.5 cm, 1.6 μm pitch, single-ended signaling
Required throughput	2 Gbps	
Probability of success	28.4%	100.0%
Latency	0.783 ns	1.185 ns
Average absolute latency error	2.76 %	4.01 %
Total power	3.466 mW	2.153 mW
Silicon area	9E-07 cm^2	44E-07 cm^2
Wire area	16E-05 cm^2	8E-05 cm^2

Table 4.7 shows that the total power dissipated by this design is 2.153 mW, which is still 38% smaller than that for the LVDS design. The design point having 5 repeaters per 0.5 cm, i.e., inserting one repeater every millimeter, can not only achieve the required throughput by dissipating lower power and occupying half the wire area compared to the LVDS circuit, but it can also guarantee an almost 100% probability of success. Though the new VSRI design results in more latency variation than the LVDS design, it is shown later in Chapter 6 that the synchronization circuit for VSRI can be designed to be insensitive to this variation. Therefore, VSRI is a very useful technique to achieve high and stable throughput at the cost of minimal power on VLSI global interconnects.

4.6 Summary

The importance of the simultaneous application of voltage scaling and repeater insertion (VSRI) to enhance the performance of low-power interconnects is discussed in this chapter. The effect of VSRI on the interconnect performance and crosstalk is analyzed in this chapter with the help of HSPICE simulation results. A comparison of VSRI with LVDS shows that it is possible to design a wave-pipelined interconnect circuit that has less power and wire area than an LVDS circuit, without any loss of throughput performance.

The effect of power supply noise, which is one of the largest sources of noise for on-chip interconnects, on VSRI is also studied in this chapter with the help of HSPICE simulations. It is shown through a comparison between VSRI and LVDS for a constant throughput constraint that unlike LVDS, VSRI overdesign can achieve the required throughput with an almost 100% probability of success in the presence of severe power supply noise.

CHAPTER 5

VOLTAGE SCALING, REPEATER INSERTION, AND WIRE SIZING OPTIMIZATION

5.1 Introduction

The simultaneous application of voltage scaling, repeater insertion, *and* wire sizing is proposed in this chapter to achieve high performance, low power, and low area on the interconnect circuits. Based on this methodology, design optimizations for three different types of applications are performed. Different design metrics are used to obtain the optimal values of supply voltage, number of repeaters, and interconnect dimensions for these applications. The design metrics are modeled based on the analytical throughput model in (2.16) and are used to pinpoint the optimal design point in a predetermined design space.

To highlight the strength and applicability of different optimal design points, these new design points are compared to each other under a constant throughput constraint. The design point obtained by optimizing the throughput-per-energy-area (TPEA) for a holistic interconnect design is also compared to low-voltage differential signaling (LVDS). The advantages of the TPEA optimization over LVDS are discussed in this chapter. Moreover, the impact of different design geometries on these

optimizations and the impact of these design optimizations on via blockage are also studied in this chapter. Finally, for latency-sensitive applications, the simultaneous application of wave-pipelining and latency-centric repeater insertion is suggested to achieve a high performance without any degradation of latency.

5.2 Voltage scaling, repeater insertion, and wire sizing

It is seen in Section 4.3 that voltage scaling is used on the interconnect circuits to reduce the total power, and repeater insertion can be used to recuperate the performance loss resulting from voltage scaling. Therefore, it is suggested in Chapter 4 that voltage scaling repeater insertion (VSRI) is a very useful technique for high-performance, low-power interconnects. However, it is seen in Section 2.3.3 that the wire dimensions also have a significant impact on the communication throughput. Altering the wire dimensions changes the resistance and capacitance of the interconnect, thereby changing its throughput. Therefore, intelligent wire sizing can also be used to recuperate the performance loss resulting from voltage scaling without significantly increasing wire area.

ITRS has projected the number of metal levels to be 12 by the end of the current decade [2]. Some wire area optimization is necessary to meet or outperform the roadmap. Optimal wire sizing can be used to achieve this because it can give a high throughput at the expense of minimal wire area.

Optimal wire sizing results in a high throughput at the expense of minimal wire area, whereas optimal repeater insertion can give a high throughput at the expense of minimal silicon area. The manufacturing cost of a VLSI system is directly related to the number of metal levels and the overall die size. A simultaneous application of voltage

scaling, repeater insertion, and wire sizing can enhance the interconnect performance while maintaining low power and low cost.

Therefore, optimal voltage scaling, repeater insertion, and wire sizing methodologies are developed for global interconnects in the following sections. The applications in three different categories are considered, and different design metrics are developed and used for different types of applications. The applications are categorized into three types as follows:

1. Low-power, high-performance applications (e.g., PDAs or cellular phones)
2. Moderate-power, moderate-performance, area-constrained applications (e.g., laptop or desktop computers)
3. Ultra-high-performance applications (e.g., liquid-cooled servers)

5.3 Design metrics for different types of applications

5.3.1 Low-power, high-performance applications

Taken to the extreme, low-power design suggests operating at the smallest possible voltage with no repeaters. However, the single interconnect throughput obtained from such design is much smaller than what is typically required for low-power applications such as PDAs or cellular phones. The number of interconnects then required to meet the throughput requirement is highly unrealistic. Therefore, to achieve the maximum throughput on low-power interconnects, a throughput-per-bit-energy (TPBE) metric is proposed. TPBE is given by the ratio of throughput to bit energy, where bit energy is the energy dissipated in the transmission of a single bit.

A metric for low-power, high-performance applications should maximize the performance and minimize the power. Logically, it seems that this can be achieved by maximizing the ratio of throughput to power. However, the power is given by the product of throughput and bit energy. As a result, the throughput term would vanish from the ratio of throughput to power. Consequently, maximizing the throughput per power would translate into minimizing only the bit energy, which fails the purpose behind this optimization. To avoid this, the throughput-independent factor of power, i.e., bit energy, is used in the construction of this metric. For a given value of power, the TPBE optimization selects a point that achieves the maximum throughput along with minimal bit energy on a single interconnect. In some ways, this metric is analogous to minimizing the energy-delay product, which is currently used as a metric in portable logic design [25].

Physically, maximizing TPBE is equivalent to minimizing the product of the total power, P_d , and the number of wires, n_w , to achieve a higher aggregate throughput, T_{req} . If T_{max} is the maximum throughput that can be achieved on a certain single interconnect, the number of parallel interconnects, n_w , required to obtain T_{req} is given by

$$n_w = \frac{T_{req}}{T_{max}} . \quad (5.1)$$

The product of P_d and n_w can be simplified to

$$P_d \cdot n_w = (E_{bit} \cdot T_{req}) \left(\frac{T_{req}}{T_{max}} \right) = \frac{T_{req}^2}{\left(\frac{T_{max}}{E_{bit}} \right)} = \frac{T_{req}^2}{TPBE} . \quad (5.2)$$

Because T_{req} is constant, (5.2) shows that the $P_d \cdot n_w$ product can be minimized by maximizing TPBE. It is also seen from (5.2) that though the TPBE metric primarily optimizes power and performance, the TPBE optimization also prevents the number of wires from increasing to an unrealistically large value.

It should be noted that the replication of the single-channel optimal design on multiple wires to achieve a higher aggregate throughput does not necessarily translate into the lowest power for that aggregate throughput. The lowest-power design for a given aggregate throughput suggests using the lowest possible supply voltage and no repeaters; however, such a design requires an unrealistically large number of parallel wires to meet the value of aggregate throughput. On the other hand, the optimal TPBE design assures that each interconnect operates at the maximum possible speed for a given bit-transition energy, thereby guaranteeing the best utilization of resources and avoiding unnecessary wire parallelism, as seen by (5.2). Multichannel optimizations, which also include an area optimization in addition to the power, are discussed in detail in the next subsection.

The main factors contributing to the bit energy are switching activity on the interconnect, subthreshold leakage, and short-circuit current. However, the results for the 180 nm technology node show that the contribution of subthreshold leakage to the total energy is less than 1% and that of energy resulting from the short-circuit current is only 5-10%, which is also consistent with the results obtained in [12]. Consequently, the energy contributions of the subthreshold leakage current and the short-circuit current are ignored while modeling the bit energy for simplicity, however, they are accounted for in the actual calculation of TPBE as shown in figures and tables later in this subsection. Bit energy is thus modeled as the switching energy, as shown in (5.3).

$$E_{bit} = \frac{1}{2} (C + n \cdot 2C_i) V_{dd}^2, \quad (5.3)$$

where $2C_i$ is the equivalent capacitance of a repeater. (Because a repeater comprises two inverters, its total capacitance C_{rep} is approximately equal to $2C_i$.) Using (2.16) and (5.3), TPBE can be expressed as

$$TPBE = \frac{1}{\left\{ \sigma_{RCseg} \ln \left(\frac{K_1}{1-v_1} \right) + \Delta_{repeater} \right\} \left\{ \frac{1}{2} (C + 2nC_t) V_{dd}^2 \right\}}. \quad (5.4)$$

TPBE can be used to determine the optimal design parameters for an interconnect circuit. To determine the optimal supply voltage, TPBE is differentiated w.r.t. the supply voltage. Setting the result equal to zero and using the expression of R_t in [6] leads to

$$\frac{V_{dd}}{2(V_{dd} - |V_t|)} = 1 + \frac{(R_{seg} C_t + 0.4 R_{seg} C_{seg}) \beta (V_{dd} - |V_t|) \ln \left(\frac{K_1}{1-v_1} \right)}{\ln \left(\frac{K_1}{1-v_1} \right) (C_t + C_{seg}) + 0.693 C_t}, \quad (5.5)$$

where V_t is the threshold voltage. The term β is given by the expressions in (2.53) and (2.54). Assuming $K_1 \sim 1$ and solving (5.5), the expression for the optimal supply voltage is obtained as

$$V_{dd,opt} = |V_t| - \left(\frac{1 - \sqrt{1 + 8\theta |V_t|}}{4\theta} \right), \quad (5.6)$$

where

$$\theta = \frac{(R_{seg} C_t + 0.4 R_{seg} C_{seg}) \beta \ln \left(\frac{1}{1-v_1} \right)}{\ln \left(\frac{1}{1-v_1} \right) (C_t + C_{seg}) + 0.693 C_t}. \quad (5.7)$$

The values of θ and the corresponding values of $V_{dd,opt}$ are shown in Table 5.1 for different values of the number of repeaters. The term $|V_t|$ is assumed to be 0.42 V [48]. The 180 nm interconnect used to calculate the values shown in Table 5.1 is 1 cm long and has cross-sectional dimensions of 250 nm x 250 nm.

Table 5.1: Values of θ and corresponding $V_{dd,opt}$.

n	θ	$V_{dd,opt}$
1	11.28	0.53 V
5	2.21	0.64 V
10	1.30	0.67 V
20	0.73	0.72 V

It is seen in Table 5.1 that θ drops as the number of repeaters increases, which results in an increase in $V_{dd,opt}$. Subsequently, the non-simplistic expression for $V_{dd,opt}$ in (5.6) results in different values of $V_{dd,opt}$ for different number of repeaters. Assuming that the number of repeaters, n , is large, (5.6) can be further simplified to

$$\frac{V_{dd}}{2(V_{dd} - |V_t|)} = 1 + X, \quad (5.8)$$

where

$$X = \theta(V_{dd} - |V_t|). \quad (5.9)$$

The values of X can be calculated using the values of θ in Table 5.1. It is observed that though the term represented by X is slightly greater than unity for a single-driver interconnect, it is significantly smaller than unity for five or more repeaters. This fact considerably simplifies (5.8) and leads to a first-order approximation for the optimal supply voltage as

$$V_{dd,opt} \approx 2|V_t|. \quad (5.10)$$

For a threshold voltage of 0.42 V (as specified for 180 nm technology [48]), the value of the optimal supply voltage is 0.84 V. However, it is seen in Figure 4.2 and Figure 5.1 that a supply voltage of 1 V results in significantly higher throughput and TPBE than that of 0.84 V. Therefore, the optimal supply voltage is rounded to 1 V. Equation (5.10), nonetheless, provides a simple first-order approximation for the optimal supply voltage.

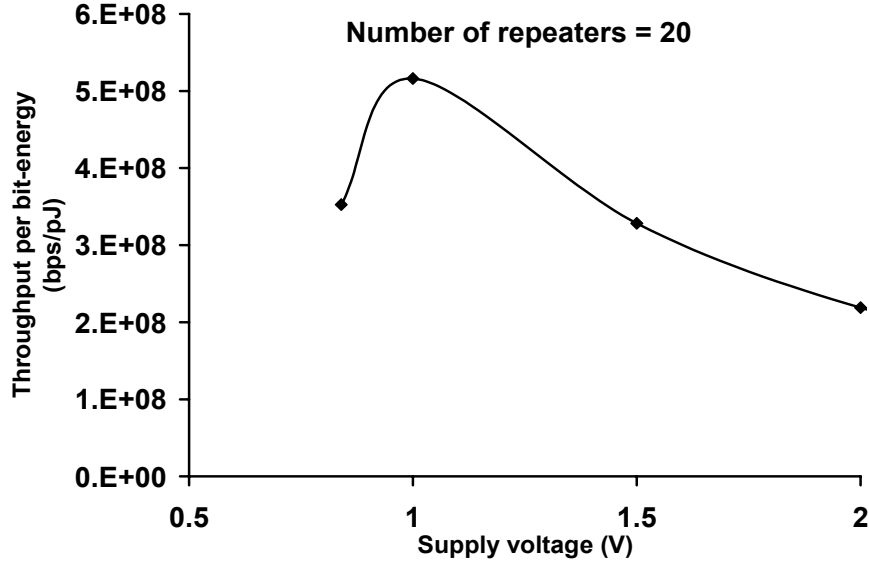


Figure 5.1: Variation of TPBE with supply voltage.

For the foundry-driven ASIC applications, interconnect dimensions are predetermined for a tier, and the parameters such as r and c are already known. In such a case, a simple analytical expression can be derived for the number of repeaters, n , that maximizes the TPBE. Setting the partial derivative of (5.4) w.r.t. n equal to zero, the expression for the optimal number of repeaters per unit interconnect length is obtained as

$$\frac{n_{opt}}{l} \approx \sqrt{\frac{(R_t c + C_t r) c + 0.8 r c C_t}{2 R_t C_t^2}}. \quad (5.11)$$

Using (5.11), the optimal number of repeaters for a 1 V supply on a 1 cm long interconnect having a square cross-section of 250 nm width is found to be 21, which is in close agreement with the value 18 obtained from HSPICE simulations shown in Figure 5.2. Therefore, for predetermined interconnect dimensions, two simple expressions given in (5.10) and (5.11) are sufficient to determine the optimal design configuration.

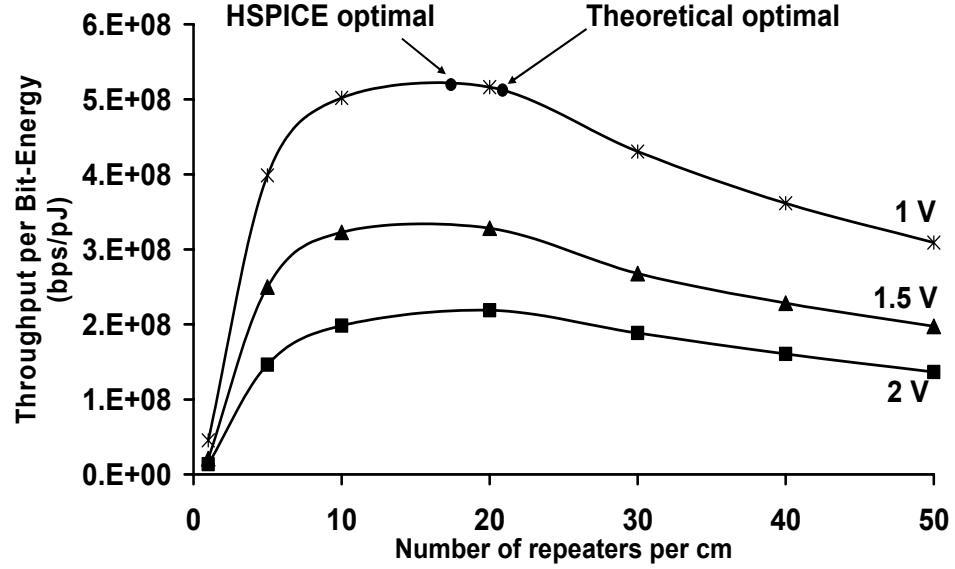


Figure 5.2: Variation of TPBE with number of repeaters on an interconnect with a 250 nm square cross-section, for different supply voltages.

The optimal values seen in Figure 5.2 are also consistent with the analysis of throughput in Figure 4.2. It is seen in Figure 4.2 for a 1 V supply that when the number of repeaters increases beyond 20, the throughput starts to saturate. However, the bit energy continues to increase, which reduces TPBE. The optimal values for TPBE thus occur just before the saturation of throughput.

For the optimal-sized repeaters, substituting the value of h_{opt} from (2.41), (5.11) simplifies to

$$\frac{n_{opt}}{l} \approx 1.183 \sqrt{\frac{rc}{R_0 C_0}}. \quad (5.12)$$

Equation (5.12) is similar to Bakoglu's expression for the optimal number of repeaters in [6], but the factor 1.183 in (5.12) differs from that in Bakoglu's expression (0.659).

Equation (5.12) suggests *inserting more repeaters* on a global interconnect because it minimizes the TPBE, whereas Bakoglu's expression minimizes the propagation delay. Substituting (5.12) in the analytical throughput expression in (2.16) and assuming $K_I \sim 1$, (2.16) can be further simplified as

$$T_{TPBE} \approx \frac{1}{\left[2.97 \ln \left(\frac{1}{1 - \nu_1} \right) + 0.693 \right] R_0 C_0}. \quad (5.13)$$

It can be seen from (5.13) and (2.30) that

$$T_{TPBE} \approx \frac{2}{3} T_{sat}. \quad (5.14)$$

It can be seen from Figure 4.2 that the optimal values of throughput given by (5.13) lie on the knee of the throughput curve, which is the region of interest.

Using the optimal supply voltage from (5.10), the optimal number of repeaters per unit length from (5.12), and using $2C_t$ as the capacitance of a repeater, the bit energy can be simplified as

$$E_{bit,opt} \approx 6.8cIV_t^2. \quad (5.15)$$

The strength of the analytical throughput expression in (2.16) lies in its ability to provide a quick estimation of many design parameters for the optimal design point. Based on this TPBE optimization, the expressions for the optimal values of various parameters are summarized in Table 5.2.

Table 5.2: Summary of design parameters for the optimal TPBE design point.

Parameter	Expression for optimal value
Supply voltage	$V_{dd,opt} \approx 2 V_t $
Repeater size	$h_{opt} = \sqrt{\frac{R_0 c}{C_0 r}}$
Optimal number of repeaters per unit interconnect length	$\frac{n_{opt}}{l} \approx 1.183 \sqrt{\frac{r c}{R_0 C_0}}$
Throughput	$T_{TPBE} \approx \frac{1}{\left[2.97 \ln \left(\frac{1}{1-v_1} \right) + 0.693 \right] R_0 C_0} \approx \frac{2}{3} T_{sat}$
Bit energy	$E_{bit,opt} \sim 6.8 c l V_t^2$

Unlike the foundry-driven ASIC designs, the interconnect dimensions in full-custom designs are determined based on the application requirements. Therefore, for full-custom designs, wire-sizing optimization can be carried out for the optimal supply voltage of 1 V to find the optimal combination of the interconnect dimensions and number of repeaters to achieve maximum TPBE. The interconnect dimensions that can be optimized are width w , height h , spacing between two interconnects s , and thickness of the dielectric t . These dimensions are shown earlier in Figure 2.3, which is redrawn as Figure 5.3 for convenience. Pitch p represents the sum ($w+s$). When all of w , h , s , and t are variable, it is not possible to obtain simple closed-form expressions for the optimal values of the individual dimensions. Instead, TPBE is calculated for all the possible combinations in a certain design space, and the design that results in maximum TPBE is selected. The energy contributions of subthreshold leakage and short-circuit current are included in this optimization analysis using the models in [12].

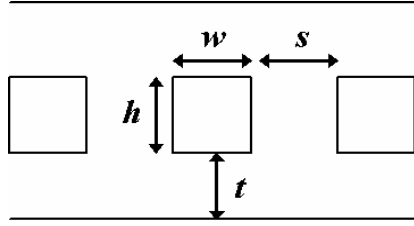


Figure 5.3: Interconnect dimensions.

The interconnect width is varied from 0.1 μm to 1 μm , the spacing-to-width ratio is varied from 1 to 5, and the interconnect height-to-width aspect ratio is varied from 1 to 2.5 based on [29]. The dielectric thickness t is assumed to be equal to the metal height h , to meet the constraints on the via aspect ratio [29]. The 1 cm long interconnect is assumed to be placed between two co-planar interconnects and two orthogonal routing planes, as shown in Figure 5.3, and the interconnect parasitics are extracted using RAPHAEL. The repeater density on the interconnect is varied from 1 to 50 repeaters per cm. In this design space, the design point that results in maximum TPBE is summarized in Table 5.3. The results in Table 5.3 show that the optimal TPBE design has a large height-to-width aspect ratio and a large spacing-to-width aspect ratio, which decrease the resistance and capacitance of the interconnect, respectively. The optimal TPBE design suggests that the interconnects should be spread out as much as possible as allowed by wire-area constraints.

Table 5.3: Optimal TPBE design point.

Parameter	Value
Interconnect length	1 cm
Supply voltage, repeater density	1 V, 5 repeaters/cm
Interconnect dimensions in μm	$w = 1, h = 2.5, s = 5, t = 2.5$
Throughput	2.004 Gbps
Total bit energy	1.468 pJ
Total power	2.942 mW
Wire area	6E-04 cm^2
Silicon area	4.45E-06 cm^2
50% Latency	1.95 ns
TPBE	1.365E+09 bps/pJ

If the optimal TPBE design in Table 5.3 is compared to a unity aspect ratio design (i.e., $w = h = s = t = 1 \mu\text{m}$) that operates with a 1 V supply and 5 repeaters per cm, it is seen that this unity aspect ratio design reduces wire area to a third and leaves silicon area unchanged. However, this design reduces throughput by 25% compared to the optimal TPBE design and it also increases bit energy by 25%. Therefore, though the unity aspect ratio design is more area-optimal, it is not suitable for low-power, high-performance applications such as PDAs and cellular phones. This comparison highlights the importance of the TPBE metric, which is specifically designed for such applications.

5.3.2 Moderate-power, moderate-performance, area-constrained applications

Parallel wire channel systems involving multiple interconnects fall under this second category of applications. For such applications, in addition to the performance, both power and area are important because they both affect the total manufacturing cost. The cost function for such systems can be considered to be the product of power and area. The total area of the system can be given as

$$A_{total} = n_w (A_{int} + A_{silicon}), \quad (5.16)$$

where A_{int} and $A_{silicon}$ are the areas occupied by the interconnect and repeaters, respectively. The cost function, i.e., the product of power and area, is therefore given by

$$P_d \cdot A_{total} = \{E_{bit} T_{req}\} \{n_w (A_{int} + A_{silicon})\}. \quad (5.17)$$

Substituting for n_w from (5.1),

$$P_d \cdot A_{total} = T_{req}^2 \frac{E_{bit} (A_{int} + A_{silicon})}{T_{max}}. \quad (5.18)$$



In (5.18), T_{req} is the constrained throughput and is therefore constant. As a result, the other term in (5.18) needs to be minimized to minimize the cost function, or in other words, the inverse of this term, i.e., throughput-per-energy-area (TPEA), needs to be maximized. The TPEA metric is thus given as

$$\text{TPEA} = \frac{T_{\max}}{E_{\text{bit}}(A_{\text{int}} + A_{\text{silicon}})} . \quad (5.19)$$

In addition to the power and performance that are optimized by a TPBE metric, TPEA also tries to optimize the total area. The maximization of TPEA for an interconnect circuit results in the maximum communication throughput along with the minimal area and power. An inspection of (5.17) and (5.19) shows that minimizing the cost function for a parallel wire channel system translates into the single interconnect TPEA optimization, which is independent of the value of constrained throughput T_{req} . In other words, *TPEA cost-performance optimization for a single interconnect also optimizes the entire parallel wire channel system*. Therefore, TPEA can be used to determine the optimal configuration for a single interconnect of a particular technology generation, and this configuration can be replicated on the required number of interconnects of that technology generation to achieve a higher aggregate throughput.

For instance, two different design choices to achieve an aggregate throughput of 6 Gbps are considered. Design I achieves a 1.5 Gbps throughput on the single interconnect, which is replicated on four wires, whereas Design II achieves a 3 Gbps throughput on the single interconnect, which is replicated on two wires. The two design configurations and the corresponding parameters are shown in Table 5.4.

Table 5.4: Two design choices to achieve an aggregate throughput of 6 Gbps.

Parameter	Design I	Design II
Aggregate throughput	6 Gbps	
Individual throughput and number of interconnects	1.5 Gbps x 4	3 Gbps x 2
Interconnect width and spacing	$w = 0.6 \mu\text{m}, s = 1.8 \mu\text{m}$	$w = 0.4 \mu\text{m}, s = 0.8 \mu\text{m}$
Interconnect layout		
Supply voltage and number of repeaters for 1 cm interconnect	1 V, 5 repeaters	1 V, 30 repeaters
Bit energy	1.71 pJ	5.30 pJ
Total power	10.26 mW	31.8 mW
Total area	$9.78\text{E-}04 \text{ cm}^2$	$2.92\text{E-}04 \text{ cm}^2$
Power-area product for 6 Gbps	$10.00\text{E-}06 \text{ W-cm}^2$	$9.28\text{E-}06 \text{ W-cm}^2$
TPEA	$3.633\text{E+}12 \text{ bps/pJ-cm}^2$	$3.797\text{E+}12 \text{ bps/pJ-cm}^2$
TPBE	$8.772\text{E+}08 \text{ bps/pJ}$	$5.660\text{E+}08 \text{ bps/pJ}$

It is seen in Table 5.4 that Design I dissipates almost a third of power compared to Design II because of its lower bit energy. However, it also requires more than three times wire area than Design II. As a result, the product of the power and area is lower in case of Design II than Design I for a constant aggregate throughput of 6 Gbps, which makes Design II a balanced design choice. However, if low-power operation is desired, Table 5.4 shows that the TPBE metric, which is introduced in the previous subsection, should be used. The TPBE optimization would result in a substantial power reduction at the expense of wire area, as seen in Table 5.4.

However, the inferences for the TPEA optimization can also be drawn without performing the entire analysis shown in Table 5.4. Table 5.4 shows that TPEA

corresponding to a single interconnect in Design II is higher than that of Design I, which makes Design II a better design choice for moderate-power, moderate-performance, area-constrained applications. The information about the throughput, bit energy, and area of a single interconnect is sufficient to compare it with the other, and it is not necessary to calculate the total power, area, and their product. When the required throughput increases, even though the total power and area scale in proportion, TPEA remains unchanged. Thus, for a given aggregate throughput, the TPEA optimization helps quickly estimate the optimal design point by achieving the balance between power and area.

The analytical throughput model in (2.16) can be used to obtain the optimal TPEA design configuration. Using (2.16) and (5.19), TPEA can be written as

$$\text{TPEA} = \frac{1}{\left(\sigma_{RCseg} \ln \left(\frac{K_1}{1-v_1} \right) + \Delta_{repeater} \right) \left(\frac{1}{2} (C + 2nC_t) V_{dd}^2 \right) (A_{int} + A_{silicon})}. \quad (5.20)$$

The partial differentiation of TPEA w.r.t. V_{dd} leads to the same expression for the optimal supply voltage as (5.10). Like in the case of TPBE, the optimal supply voltage is rounded to 1 V. For the foundry-driven ASIC applications with predetermined interconnect dimensions, partial differentiation of TPEA w.r.t. n results in the following equation, which can be numerically solved.

$$C \left(\frac{R_t C A_{int}}{n^2} + \frac{1.8 R C_t A_{int}}{n^2} + \frac{0.4 R C A_{rep}}{n^2} + \frac{0.8 R C A_{int}}{n^3} - 3 R_t C_t A_{rep} \right) = 2 C_t^2 \left(2 n R_t A_{rep} + R_t A_{int} + R A_{rep} \right), \quad (5.21)$$

where A_{rep} is the area of a single repeater, i.e., $A_{silicon} = n A_{rep}$. For a 1 V supply, the optimal number of repeaters for a 1 cm long interconnect having a square cross-section

with 250 nm width is found to be 12, which is in close agreement with the number 10 obtained from HSPICE simulations shown in Figure 5.4.

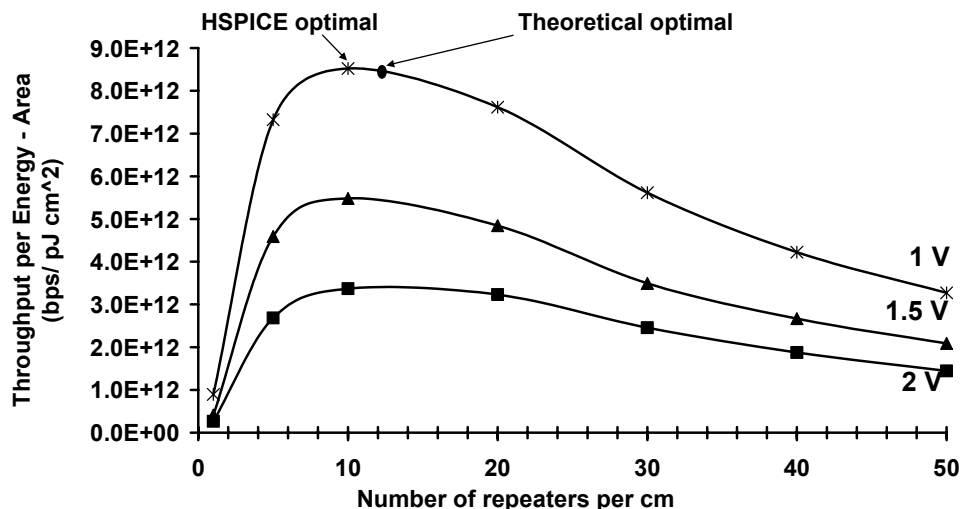


Figure 5.4: Variation of TPEA with number of repeaters on an interconnect with 250 nm square cross-section, for different supply voltages.

For the full-custom applications, the interconnect dimensions are varied in the same manner as the TPBE analysis to perform wire-sizing optimization. The resulting optimal TPEA design point is summarized in Table 5.5. A comparison of Table 5.3 and Table 5.5 shows that because the TPEA metric optimizes area in addition to power, the optimal TPEA design suggests using smaller values of w and s .

Table 5.5: Optimal TPEA design point.

Parameter	Value
Interconnect length	1 cm
Supply voltage, repeater density	1 V, 6 repeaters/cm
Interconnect dimensions in μm	$w = 0.2, h = 0.5, s = 0.4, t = 0.5$
Throughput	1.66 Gbps
Total bit energy	1.86 pJ
Total power	3.088 mW
Wire area	$6\text{E-}05 \text{ cm}^2$
Silicon area	$5.33\text{E-}06 \text{ cm}^2$
50% Latency	3.64 ns
TPEA	$1.366\text{E+}13 \text{ bps/pJ cm}^2$

For this type of applications, the primary advantage of the TPEA optimization is that a single interconnect optimization also optimizes the parallel wire channel design for any throughput constraint. Additionally, the simplicity and the holistic nature of this design optimization and its scalability to interconnects of any lengths in any technology generation are some of its other advantages.

5.3.3 Ultra-high-performance applications

Having the highest possible repeater density or the largest interconnect dimensions may seem necessary for achieving ultra-high performance for this category of applications. However, it is seen from Figure 5.5 that when the throughput enters the saturation regime, increasing the repeater density significantly increases the silicon area without any noticeable enhancement of the throughput. Similarly, for a constant wire pitch, increasing the wire width ceases to increase the throughput beyond a certain limit as a result of the high mutual capacitance. If the wire pitch is not constant, i.e., the width and the spacing increase in the same proportion, the throughput increases because R reduces and C remains unchanged. However, this increase in the throughput is achieved at the expense of a significant increase in the wire area, which is not desirable for the processor architectures that are already wire-limited [34].

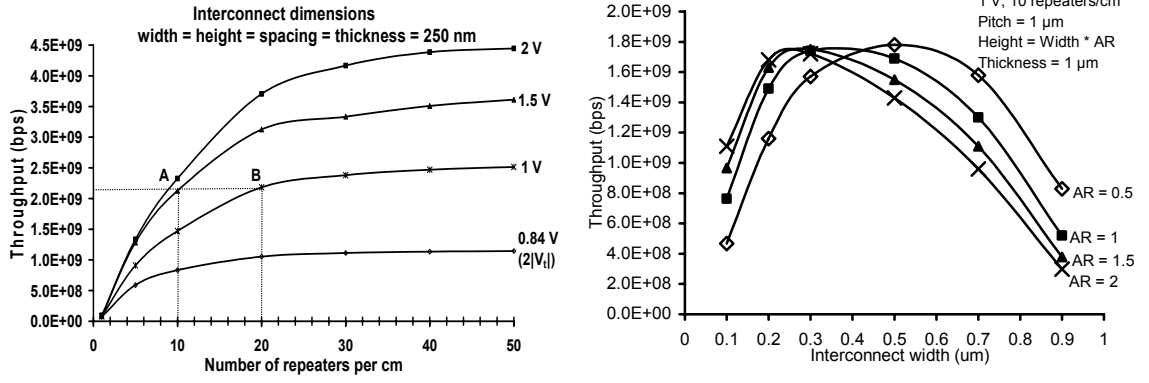


Figure 5.5: Effect of repeater density and wire dimensions on throughput performance.

Therefore, to obtain the optimal combination of the interconnect dimensions and number of repeaters to achieve very high throughput performance, a throughput-per-area (TPA) metric can be used, in which area is given by the sum of silicon area and wire area. The maximization of TPA helps achieve a high throughput on the interconnect circuit at the expense of minimal area. A similar metric has been proposed in [45] for the wire-width optimization of interconnects having square cross-sections. However, the TPA metric proposed in this research includes the optimization of silicon area in addition to the optimization of *all* interconnect dimensions. Using (2.16), TPA can be written as

$$TPA = \frac{T_{\max}}{A_{\text{int}} + A_{\text{silicon}}} = \frac{1}{\left\{ \sigma_{RCseg} \ln \left(\frac{K_1}{1 - \nu_1} \right) + \Delta_{\text{repeater}} \right\} \{ A_{\text{int}} + nA_{\text{rep}} \}}. \quad (5.22)$$

Physically, maximizing TPA is equivalent to minimizing the total area of a parallel wire system that uses n_w wires to achieve a higher aggregate throughput, T_{req} . Similar to (5.16), the total area is given by

$$A_{\text{total}} = n_w (A_{\text{int}} + A_{\text{silicon}}) = \frac{T_{\text{req}}}{T_{\max}} (A_{\text{int}} + A_{\text{silicon}}) = \frac{T_{\text{req}}}{TPA}. \quad (5.23)$$

Equation (5.23) shows that because T_{req} is constant, the total area of a parallel wire system is minimized through the maximization of TPA.

An inspection of Figure 4.2 shows that the throughput drops with the decrease in the supply voltage. Therefore, voltage scaling is not very useful for ultra-high-performance applications. The supply voltage of 1.8 V - 2 V, which is typically used for 180 nm interconnect circuits, is the optimal supply voltage in this case.

For the foundry-driven ASIC applications with predetermined interconnect dimensions, the partial differentiation of (5.22) w.r.t. n results in the equation

$$\left(R_t C_t + \frac{R_t C}{n} + \frac{C_t R}{n} + 0.4 \frac{RC}{n^2} \right) A_{rep} = \left(\frac{R_t C}{n^2} + \frac{C_t R}{n^2} + \frac{0.8 RC}{n^3} \right) (n A_{rep} + A_{int}), \quad (5.24)$$

which can be numerically solved to obtain the optimal number of repeaters. For a 1 cm long interconnect having a square cross-section of 250 nm width and operating with a supply voltage of 2 V, this number is found to be 28, which is in close agreement with the number 25 obtained from HSPICE simulation results shown in Figure 5.6.

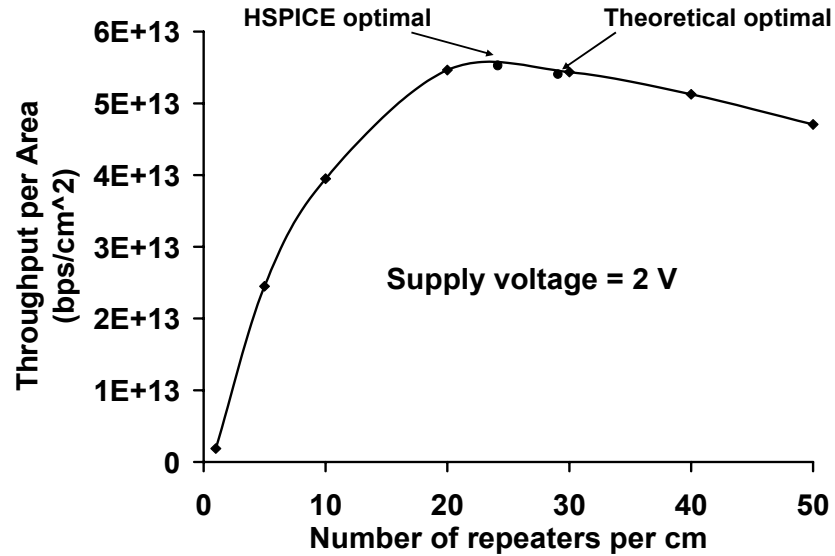


Figure 5.6: Variation of TPA with number of repeaters on an interconnect with 250 nm square cross-section, for a 2 V supply.

As in the case of TPBE and TPEA, the interconnect dimensions can also be optimized using TPA for full-custom applications, and the resulting optimal design point is summarized in Table 5.6. It is seen in Table 5.6 that the optimal TPA design suggests using relatively small values of w and s and a large repeater density. It should be noted that TPA for a parallel wire channel system can be obtained by multiplying both the numerator and the denominator of (5.22) by n_w and is essentially the same. Thus, single channel TPA optimization also optimizes the entire interconnect system for ultra-high-performance applications.

Table 5.6: Optimal TPA design point.

Parameter	Value
Interconnect length	1 cm
Supply voltage, repeater density	2 V, 30 repeaters/cm
Interconnect dimensions in μm	$w = 0.2, h = 0.5, s = 0.2, t = 0.5$
Throughput	4.89 Gbps
Total bit energy	23.68 pJ
Total power	115.79 mW
Wire area	4E-05 cm^2
Silicon area	2.67 E-05 cm^2
50% Latency	3.49 ns
TPA	7.331E+13 bps/ cm^2

5.4 Comparison of optimal design points

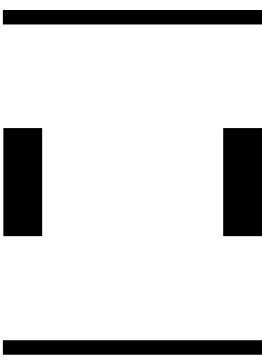


Different approaches to the interconnect optimization for different types of applications are discussed in the previous section. It is seen in Section 5.3.1 that the optimal TPBE design, which does not include any area optimization, suggests the use of relatively thicker interconnects. However, for the area-sensitive applications, the use of densely-packed thinner interconnects is suggested in Section 5.3.3. Similarly, for the power-centric applications, using a lower supply voltage and a smaller repeater density is

recommended, whereas for the optimal TPA design, which does not focus on reducing power, the use of a high supply voltage along with a high repeater density is suggested. The three optimal design points are summarized in Table 5.7 for convenience. The interconnect geometries corresponding to these optimal design points are presented in Table 5.8. The interconnects in Table 5.8 are drawn according to their true proportions.

Table 5.7: Optimal design points for different applications.

Type of application	Metric used	Optimal design point (supply voltage, repeater density, wire cross-sectional dimensions, spacing)
Low power, high performance	Throughput-per-bit-energy (TPBE)	1 V, 5 repeaters/cm 1 μm x 2.5 μm , $s = 5 \mu\text{m}$
Moderate power, high performance, constrained area	Throughput-per-energy-area (TPEA)	1 V, 6 repeaters/cm 0.2 μm x 0.5 μm , $s = 0.4 \mu\text{m}$
Ultra-high performance	Throughput-per-area (TPA)	2 V, 30 repeaters/cm 0.2 μm x 0.5 μm , $s = 0.2 \mu\text{m}$

Table 5.8: Interconnect geometries and other parameters corresponding to different optimal design points.

Design optimization	Optimal TPBE	Optimal TPEA	Optimal TPA
Configuration	1 V, 5 repeaters/cm	1 V, 6 repeaters/cm	2 V, 30 repeaters/cm
Interconnect geometry			
Throughput (Gbps)	2.004	1.660	4.890
50% Latency (ns)	1.95	3.64	3.49
Total power (mW)	2.942	3.088	115.79
Wire area (cm^2)	60E-05	6E-05	4E-05
Silicon area (cm^2)	4.45E-06	5.33E-06	26.70E-06

It is interesting to see how these optimal design points compare with a latency-centric design. Therefore, in the same design space that is used for the optimizations in Section 5.3, the design point that results in a minimal latency is chosen. This optimal latency-centric design point is summarized in Table 5.9. The values of TPBE, TPEA, TPA, and latency in the entire design space are shown in Appendix C.

Table 5.9: Optimal latency-centric design point.

Parameter	Value
Interconnect length	1 cm
Supply voltage, repeater density	2 V, 1 repeater/cm
Interconnect dimensions in μm	$w = 1, h = 2.5, s = 5, t = 2.5$
Throughput	1.25 Gbps
Total bit energy	3.62 pJ
Total power	4.53 mW
Wire area	$6\text{E-}04 \text{ cm}^2$
Silicon area	$8.89 \text{ E-}07 \text{ cm}^2$
Latency	0.742 ns

If the three optimal design points in Table 5.7 and the optimal latency-centric design point in Table 5.9 are considered together, it is seen that the optimal latency-centric design results in the lowest wave-pipelined throughput of 1.25 Gbps among these designs. Therefore, these four designs are compared in terms of power and area for a constant single channel throughput of 1.25 Gbps, as shown in Table 5.10, for a clearer understanding of the design trade-offs. It is evident from Table 5.10 that among these designs, the optimal TPBE design results in the minimal power for this performance, whereas the optimal TPA design minimizes the wire area. Though the optimal TPA design results in larger silicon area than the other designs, it prevents the silicon area from increasing to a value where it would be redundant. Moreover, the optimal TPA design, which is capable of achieving a 4.89 Gbps throughput using this silicon area, is

operated at a 1.25 Gbps throughput in this analysis, and is therefore significantly underutilized.

Table 5.10: Comparison of optimal design points for a constant throughput of 1.25 Gbps.

Parameter	Optimal TPBE design	Optimal TPEA design	Optimal TPA design	Optimal latency Design
Throughput	1.25 Gbps			
Total power	1.83 mW	2.33 mW	29.62 mW	4.53 mW
Wire area	60E-05 cm ²	6E-05 cm ²	4E-05 cm ²	60E-05 cm ²
Silicon area	4.45E-06 cm ²	5.33E-06 cm ²	26.71E-06 cm ²	0.89E-06 cm ²

It is seen from Table 5.10 that the optimal TPEA design does not result in the lowest power or area among these designs, but it achieves a balance between the two and keeps both power and area close to their minimum possible values. Therefore, the TPEA optimization is truly a holistic design choice.

It is also seen in Table 5.10 that the optimal latency-centric design results in considerably high values of power and wire area for the given throughput performance. Though the latency-centric design minimizes the silicon area, it should be noted that this design also results in a minimal throughput. Unlike the latency-centric design, the other three designs are capable of achieving a higher throughput without any changes to their existing configurations. If the latency-centric design approach is used, the maximum single interconnect throughput in the chosen design space, which includes almost all feasible design choices for the 180 nm node, is restricted to a small value of 1.25 Gbps.

5.5 Comparison of LVDS and TPEA optimization

As discussed in the previous section, TPEA balances performance, power, and area for single as well as parallel wire channel systems, and it can be efficiently used in

the design of on-chip bus networks. A specific example of one such bus is a 32-byte L2 cache bus used in the fifth-generation SPARC64 processor that achieves 41.6 GB/s bandwidth at 1.3 GHz clock [67]. As seen in Table 5.5, the optimal TPEA design, which achieves a maximum throughput of 1.66 Gbps, can be used to design this cache bus because the bus requires a throughput of only 1.3 Gbps per wire (to operate from a 1.3 GHz clock).

Low-voltage differential signaling (LVDS) is another technique that can be used to design this cache bus. As shown in [63], a 0.5 cm LVDS interconnect having a voltage swing of 600 mV on the differential interconnect can easily meet the throughput requirement of the SPARC memory bus [67]. The two design choices for the SPARC memory bus are shown in Figure 5.7. It should be noted that the interconnect length in this case is 0.5 cm, therefore, the optimal TPEA design of 1 V and 6 repeaters/cm translates into 3 repeaters for 0.5 cm length.

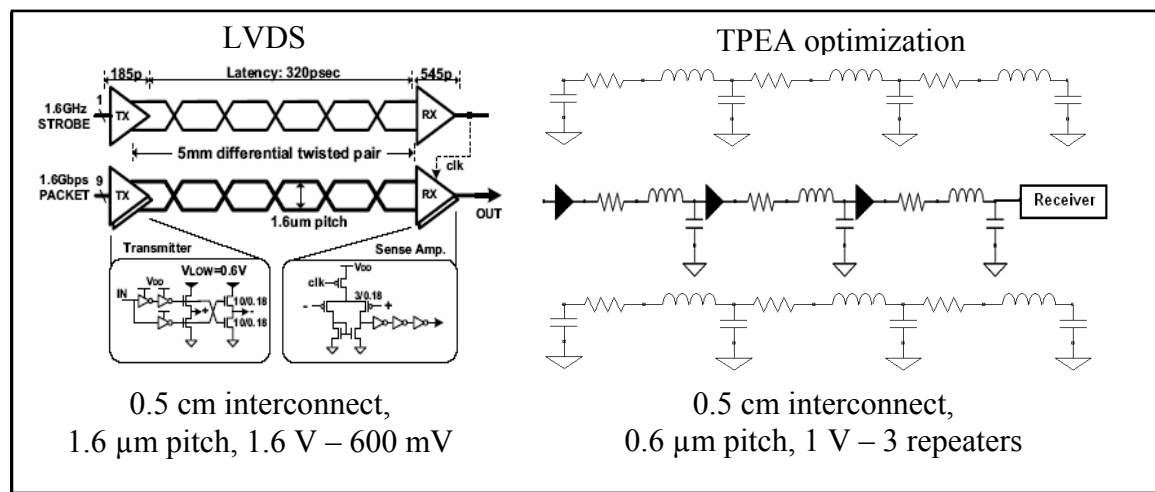


Figure 5.7: Two different design choices for 1.3 Gbps SPARC memory bus.

In the LVDS circuit, though the interconnect power dissipation is small because of the use of a lower voltage swing, the drivers and receivers are fairly large and operate from a full-swing supply voltage of 1.6 V. Therefore, the drivers and receivers dissipate a lot of power, which results in a net increase in the total power of the circuit. The LVDS circuit also occupies a lot of wire area because of its large wire pitch and differential signaling. Moreover, though this circuit easily meets the required throughput on a 0.5 cm link, interconnects as long as 1.5 cm may be required for similar applications [67], where the circuit parameters shown in Figure 5.7 cannot be used directly. To achieve the same throughput performance on longer interconnects, larger transistor or wire sizes, or a larger voltage swing may be required, which could significantly increase the total power and/or area of the LVDS circuit.

On the other hand, in the case of TPEA optimization, T_{max} is a function of the number of repeaters per unit length, and not the number of repeaters or interconnect length alone. Therefore, the optimal TPEA design configuration in Table 5.5 can be easily extended to any interconnect length by maintaining the same repeater density, to achieve the same throughput. The use of smaller wire pitch and single-ended signaling in the TPEA optimization result in significantly less wire area compared to LVDS. Taking advantage of this fact, the data lines are assumed to be shielded by co-planar ground lines in the optimal TPEA design for the SPARC memory bus, as shown in Figure 5.7, to prevent the possible reduction in throughput from dynamic delay effects.

Designs for a 0.5 cm link using the LVDS circuit in [63] and the optimal TPEA configuration in Table 5.5 are compared in Table 5.11 for a single channel throughput performance of 1.3 Gbps to achieve a bandwidth of 332.8 Gbps (41.6 GB/s) on the 32-

byte SPARC memory bus [67]. It is seen in Table 5.11 that the optimal TPEA design results in a 12% decrease in dynamic power and more than a 60% decrease in interconnect area compared to LVDS, without any loss of throughput performance.

Table 5.11: Comparison of LVDS and TPEA optimization for a 0.5 cm link.

Parameter	LVDS 1.6 V, 600 mV diff.	TPEA Optimization 1 V, 3 repeaters
Bus bandwidth	1.3 Gbps x 32 bytes = 332.8 Gbps (41.6 GB/s)	
Interconnect dimensions	1.6 μm pitch, differential interconnect	Pitch = 0.6 μm Width = 0.2 μm Height = 0.5 μm
Latency	0.92 ns	1.58 ns
* Switching energy	0.913 pJ/bit	0.809 pJ/bit
Dynamic power	303.84 mW	269.24 mW
Interconnect area	4.09E-02 cm^2	1.54E-02 cm^2
Silicon area	3.81E-04 cm^2	6.82E-04 cm^2
Power density	482 W/ cm^2	191 W/ cm^2
TPEA	8.83 Tbps/pJ. cm^2	25.58 Tbps/pJ. cm^2

* Switching energy is considered here instead of total energy because [63] provides only the switching energy values.

The optimal TPEA design results in some increase in silicon area compared to LVDS, but this increase is tolerable because the processor architectures that use these techniques are primarily wire-limited [34]. Even though the TPEA optimization results in more overhead capacitance as a result of larger silicon area compared to LVDS, it uses a scaled supply voltage (1 V instead of 1.6 V), which results in lower overall power dissipation than LVDS. This decrease in power dissipation is commendable because LVDS is a technique specially designed for low-power interconnects. Moreover, because the TPEA optimization tries to minimize the area in addition to power, a significant reduction in the wire area is also obtained for the same throughput performance.

It is important to note that the interconnect area in the case of optimal TPEA design also accounts for the area occupied by shielding ground lines. If dynamic delay effects do not considerably reduce the maximum throughput for this design point, i.e., from 1.66 Gbps to a value less than 1.3 Gbps, no shielding lines would be required, which would result in a more than 80% reduction in the wire area compared to LVDS.

Table 5.11 also shows the values for the power density, which is given by the ratio of the device power to the silicon area. For high-performance VLSI designs, maintaining low power density can be as important as maintaining low power. It is seen in Table 5.11 that the power density for the optimal TPEA design is only 40% of that of the LVDS design, which is another reason for choosing this design over LVDS.

5.6 Impact of interconnect geometries on different design optimizations

It is seen in Section 3.8.1 that inserting ground lines is an effective technique to reduce crosstalk. For global interconnects, one ground line can be inserted for one or more signal lines. Therefore, different interconnect geometries involving the ground lines are considered in this section, and the impact of these geometries on all three design optimizations in Section 5.3 is studied.

In the design D1, a ground line (G) is inserted after every five signal lines (S), whereas in the design D2, one ground line is inserted for every signal line. The two designs are shown in Figure 5.8. For the RC analysis of the signal line surrounded by two other signal lines, the worst case switching for the mutual capacitance is considered, i.e., the adjacent interconnects simultaneously switch in the opposite direction. Therefore, the

worst case throughput is accounted for in D1. For D2, the neighbors of the signal lines are quiet ground lines.

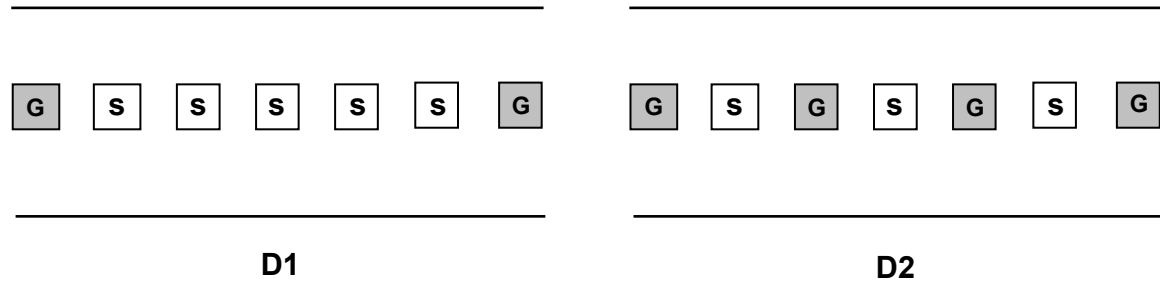


Figure 5.8: Different design geometries for an interconnect system.

All of TPBE, TPEA and TPA design optimizations are performed again to obtain the optimal design point and the optimal design geometry. The results of these new optimizations are shown in Table 5.12. The comparison between Table 5.7 and Table 5.12 shows that the optimal design points remain unchanged for all cases. However, different design optimizations suggest using different interconnect geometries.

Table 5.12: Design optimizations for two different geometries.

Type of optimization	Optimized value of metric (D1)	Optimized value of metric (D2)	Optimal geometry	Optimal design point (supply voltage, repeater density, wire cross-sectional dimensions, spacing)
TPBE	1.365E+09 bps/pJ	1.518E+09 bps/pJ	D2	1 V, 5 repeaters/cm 1 μm x 2.5 μm , $s = 5 \mu\text{m}$
TPEA	1.366E+13 bps/pJcm ²	0.505E+13 bps/pJcm ²	D1	1 V, 6 repeaters/cm 0.2 μm x 0.5 μm , $s = 0.4 \mu\text{m}$
TPA	7.331E+13 bps/cm ²	2.784E+13 bps/cm ²	D1	2 V, 30 repeaters/cm 0.2 μm x 0.5 μm , $s = 0.2 \mu\text{m}$

Though the insertion of ground lines enhances the throughput, reduces the crosstalk, and reduces the bit energy, it also results in some wire area overhead. Intuitively, the TPEA and TPA optimizations, which include wire area optimization, suggest using fewer ground lines. However, in the case of TPBE optimization, inserting more ground lines translates into a higher throughput and lower bit energy, so design D2 is preferred over D1. Therefore, it can be concluded that for the area-insensitive applications, the use of alternate ground and signal lines helps reduce power, enhance performance, and also maintain good signal integrity.

5.7 Impact of simultaneous application of voltage scaling, repeater insertion, and wire sizing on via blockage

The optimal design configurations for the throughput-centric interconnects suggest inserting more repeaters on the interconnects than that are typically inserted on the latency-centric interconnects. The increase in the number of repeaters translates into an increase in the number of vias, which could further reduce the wiring efficiency of all metal levels. However, it is shown in this section that repeater insertion, when accompanied by optimal wire sizing, can reduce the total via area, without any loss of throughput performance.

To underline this fact, the conventional latency-centric approach for the interconnect design is compared to the throughput-centric wave-pipelining approach in Table 5.13. The application chosen is the same SPARC cache bus that is used in Section 5.5. A 1 cm long metal-5 global interconnect in [29] (whose dimensions are shown in Table 1.1) that has 4 repeaters per cm, as suggested by latency-centric repeater insertion

in [6], achieves the required 1.3 Gbps throughput when operating from a 2 V supply. The same throughput can also be obtained on a metal-3 interconnect in [29] (whose dimensions are also shown in Table 1.1), which uses a smaller wire pitch, 10 repeaters per cm, and a scaled supply voltage of 1 V. Because the latter is routed on metal level 3 instead of metal level 5, it causes via blockage on fewer number of metal levels compared to the latency-centric design. Moreover, supply voltage scaling in the case of wave-pipelined design results in less power dissipation compared to the latency-centric design, without any loss of throughput performance.

Table 5.13: Comparison of latency-centric and throughput-centric design approaches in terms of via blockage, power, area.

Design approach	Latency-centric	Throughput-centric
Bandwidth	41.6 GB/s (1.3 Gbps per wire)	
Configuration	2 V, 4 repeaters/cm	1 V, 10 repeaters/cm
Wire pitch	1600 nm	600 nm
Routing level	M5	M3
Total number of repeaters	1024	2560
Power dissipation	3.25 W	0.82 W
Wire area	0.04 cm ²	0.015 cm ²
Via area	9.821E-05 cm ²	5.474E-05 cm ²

It is seen in Table 5.13 that the simultaneous application of voltage scaling, repeater insertion, and wire sizing in the throughput-centric design results in a 75% reduction in power, a 60% reduction in wire area, and more than a 40% reduction in via area compared to the latency-centric design. Though the throughput-centric approach increases repeater area, repeater area is only a small percentage of the total silicon area [60], and *an increase in the repeater area in this case actually reduces via area*. Moreover, future system architectures are primarily projected to be wire-limited [34], which indicates that they can easily tolerate this increase in silicon area.

As seen in Table 5.13, instead of using large wire widths for global interconnects, smaller-sized wires with more repeaters can be used to achieve the same throughput. The wire area saved from scaling the interconnect dimensions can then be used to insert shielding ground lines, which would provide excellent return paths for the high-frequency currents and minimize the performance variations resulting from inductive and capacitive coupling.

Intelligent wire sizing enables the elimination of upper metal layers or scaling of wires on the existing metal layers. In either case, this approach reduces the total system cost, which is one of the primary driving forces for the present VLSI system designs. Therefore, it is seen in this chapter that the simultaneous application of voltage scaling, repeater insertion, and wire sizing is an effective design technique for low-power, low-cost VLSI global interconnects. This design methodology gives the designer a large space to trade-off between performance, power, and area to design an optimal circuit to meet the application requirements.

5.8 Latency-sensitive wave-pipelining

Wave-pipelining is a throughput-centric approach, which advocates that the latency is only a one-time delay and a small increase in the latency can be tolerated if it is accompanied by a significant enhancement of the throughput. In simpler words, wave-pipelining suggests that even though the first data bit is delayed, the following data bits are received at a very high bit rate, which facilitates the data processing at a very high frequency. A cache bus is a good example of a latency-insensitive application. Because the cache bus transfers large volumes of data to and from cache memory, the *rate* at

which data are transferred is extremely important to maintain high performance. In such cases, latency is only a one-time delay and it loses its significance once the data transfer is initiated. Similarly, the links in application-specific networks-on-chip (NoC), whose data rate is decoupled from the latency, are another example of a latency-insensitive application [68]. Because the data transfer in an NoC or SoC could take multiple clock cycles, the VLSI design trend is moving toward the *latency-insensitive design* [69].

However, the latency can be very important for some high-performance applications, such as bypass buses, which forward the data from one unit to the other to resolve the pipeline stalls. For example, Intel's next-generation 64-bit Itanium microprocessor, which is manufactured using the 180 nm bulk technology, uses a 2 mm long bypass bus, which operates with low latency and high throughput [1]. When certain modules in the processor architectures are *waiting* on the data from prior instructions to generate new results or make a decision, the delay after which the first data bit arrives can be extremely important. Moreover, bypass buses may not carry large volumes of consecutive data bits, but they could carry only one set of data at a given time. Therefore, the throughput of the bus may not be as critical in these cases as is the latency.

For example, MIPS (million instructions per second) instructions are classically pipelined into five stages [70] as follows:

1. Instruction Fetch (IF)
2. Decode and register read (DEC)
3. Arithmetic logic unit (ALU) execution (EX)
4. Data memory access (DM)
5. Register write-back (WB)

Every stage in the five-stage pipeline shown in Figure 5.9 can be further split into multiple stages such that one small stage is completed every clock cycle. However, if Instruction 2 in Figure 5.9 writes the data into a certain register and Instruction 3 needs to read and process the data from the *same* register, Instruction 3 may have to be stalled until Instruction 2 finishes WB. However, the data needed by Instruction 3 is ready at the end of EX of Instruction 2, which can be directly *forwarded* to the EX input of Instruction 3 to resolve this data hazard. This is achieved by using the bypass bus shown in Figure 5.9. Bypass buses may also be used to forward the data from the DM stage of one instruction to the input of DEC or ALU stage of a following instruction.

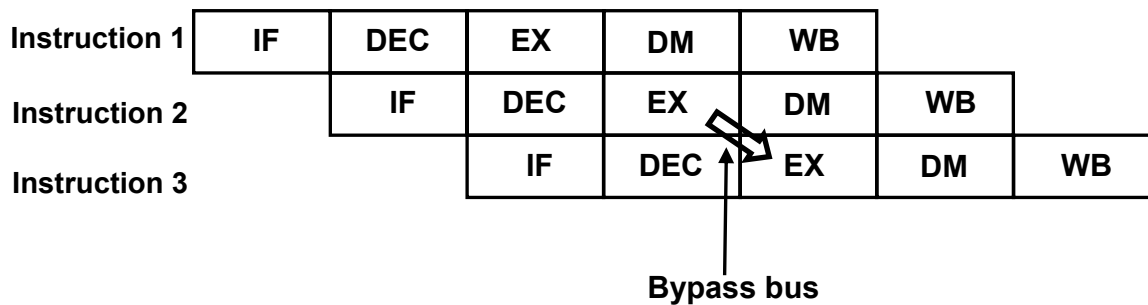


Figure 5.9: Five-stage MIPS pipeline.

A partial pipelined datapath with a bypass bus for a typical MIPS architecture is shown in Figure 5.10. Bypass buses carry the data, which are used by following instructions that depend on these data. Therefore, the sooner the data value is obtained, the sooner the dependent instructions can start executing. For such *latency-sensitive* applications, a combination of latency-centric repeater insertion and wave-pipelining is suggested in this section.

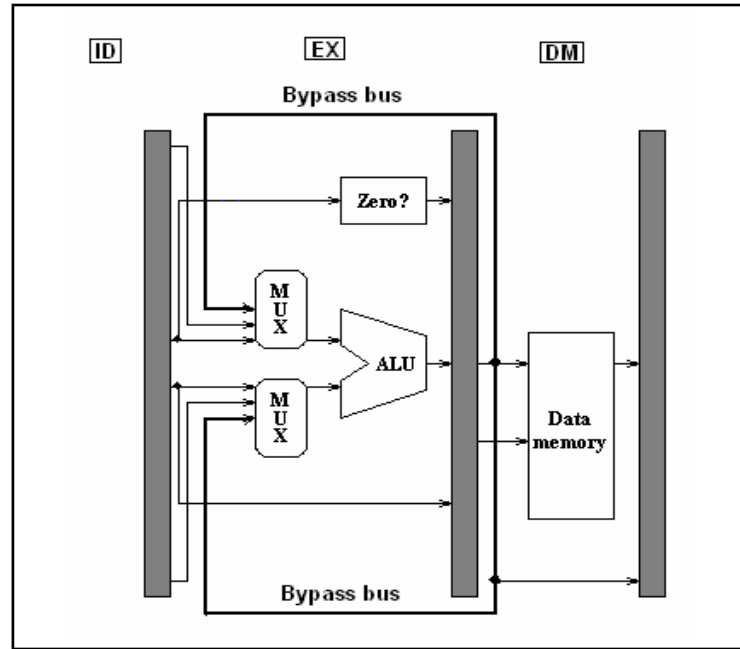


Figure 5.10: Bypass bus in pipelined partial datapath.

5.8.1 Optimal latency-centric wave-pipelining

Bakoglu and Meindl have derived the optimal number and size of repeaters to minimize the delay of an interconnect circuit [6]. Operating at this design point assures the minimal latency on the interconnect. Typically, these interconnects carry only a single data bit in one clock cycle, i.e., a new data bit is sent when the previous data bit reaches the receiver. However, without changing the design configuration, the same interconnect can be wave-pipelined, i.e., a new data bit can be sent after the previous bit achieves a certain voltage swing on the first segment (as shown in Section 2.3.1). Therefore, combining latency-centric repeater insertion and wave-pipelining not only assures the minimal latency but it also guarantees the maximum possible throughput for that design configuration.

It is important to note that scaling down the supply voltage typically increases the latency because of a lower current drive to the active circuits. As a result, voltage scaling is not recommended for the latency-sensitive applications. Because the focus is on minimizing the latency, a typical value of the supply voltage, i.e., 2 V for the 180 nm applications, is used in this analysis.

The resistance and capacitance parameters corresponding to a 1 cm long metal-5 interconnect [29] and the transistors for the 180 nm technology node are presented in Table 5.14. Using the expressions in [6], the optimal number and size of the repeaters are calculated for this interconnect circuit. Based on [6], Table 5.14 shows the value of the throughput that can be achieved by the latency-centric approach. The maximum throughput that can be obtained by sending the data in the wave-pipelined fashion, which is calculated using (2.16), is also shown in Table 5.14.

Table 5.14: Latency-centric repeater insertion and wave-pipelining.

Approach	Latency-centric repeater insertion	Throughput-centric wave-pipelining
Metal-5 interconnect parameters	$R = 172 \text{ ohm/cm}$, $L = 2.606 \text{ nH/cm}$, $C = 2.226 \text{ pF/cm}$	
Transistor parameters	$R_0 = 10080 \text{ ohm}$, $C_0 = 2.32 \text{ ff}$	
Optimal number and size of repeaters	$n_{opt} = 3 \text{ repeaters per cm}$, $h_{opt} = 237$	
Latency	0.719 ns	
Throughput	1.39 Gbps	3.42 Gbps

The latency shown in Table 5.14 is the 50% latency at the input of the receiver circuit. It is seen from Table 5.14 that the maximum throughput achieved by wave-pipelining is almost 2.5 times the reciprocal latency. Wave-pipelining thus facilitates sending the data bits with a more than double the bit rate, *without any degradation of the latency*. Thus, the combination of latency-centric repeater insertion and wave-pipelining

achieves the best of throughput and latency for an interconnect circuit and is an ideal design choice for the latency-sensitive applications.

If the bypass bus shown in Figure 5.9 needs to forward data to the input of ALU in multiple *consecutive* instructions, the optimal latency-centric repeater insertion would deliver every ALU result after a latency of 0.719 ns. However, when the same interconnect is wave-pipelined, the first set of data arrives after a 0.719 ns delay, and the following sets arrive after an additional delay of only 0.292 ns (i.e., inverse of 3.42 Gbps). Consequently, the combination of latency-centric repeater insertion and wave-pipelining facilitates the operation at a higher clock frequency. If an increase in the clock frequency is used to split one pipeline stage into multiple smaller stages, each of which takes one clock period, this approach could significantly enhance the instruction throughput of the pipeline.

5.8.2 Wave-pipelining with suboptimal repeater insertion

For the interconnect circuits that result in a flat optimal for the latency curve, more repeaters can be added to increase the throughput without considerably increasing the latency. For instance, the latency curve for a 1 cm long 180 nm interconnect that has cross-sectional dimensions of 250 nm x 250 nm is shown in Figure 5.11. The values of latency correspond to a 2 V supply. It is seen from Figure 5.11 that the minimum latency of 1.58 ns is obtained with six repeaters, which would translate into a throughput of 0.633 Gbps with the latency-centric approach or a throughput of 1.6 Gbps with wave-pipelining.

Suboptimal repeater insertion similar to [53] would suggest operating with four repeaters to obtain a latency of 1.65 ns, which is only 4% more than the minimum latency. However, a careful observation of Figure 5.11 shows that because of the flatness of the

curve near its minima, a latency of 1.65 ns can also be obtained with eight repeaters. If this interconnect is wave-pipelined using eight repeaters instead of four or six, a 2 Gbps throughput is obtained, with a negligible change in the latency.

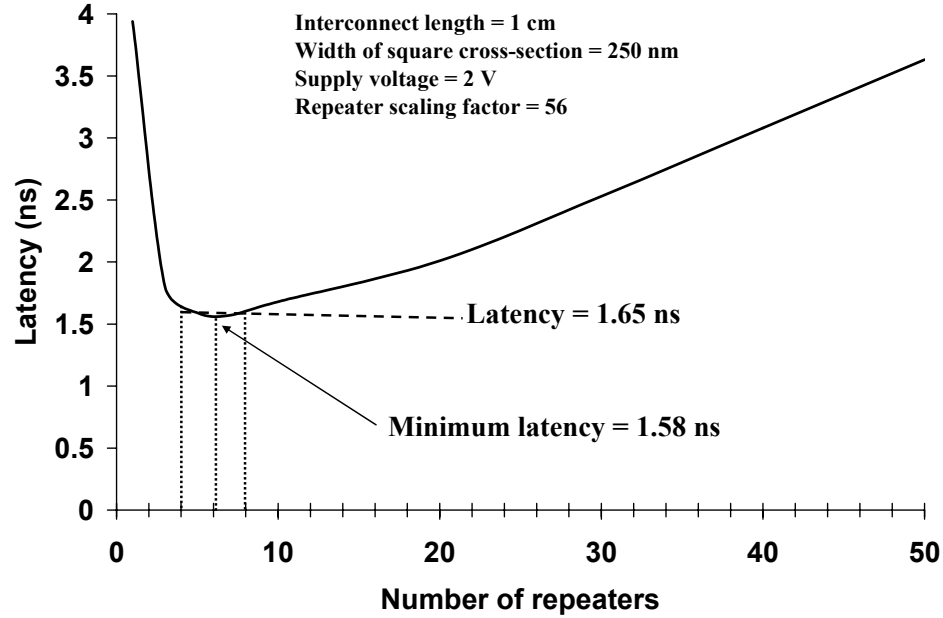


Figure 5.11: Flat minima for latency of a 180 nm interconnect.

The values of throughput for these three design points are shown in Figure 5.12. *It is seen in Figure 5.12 that if the interconnect is wave-pipelined using eight repeaters, it results in only a 4% increase in latency, but it increases the throughput by more than three times compared to the optimal reciprocal latency.* As shown in Figure 5.12, the throughput obtained with eight repeaters is significantly higher than that obtained with four or six repeaters. Therefore, if the latency curve for an interconnect circuit has a flat minima, a few more repeaters can be inserted than that suggested by the optimal latency-centric design to significantly enhance the throughput, *without any noticeable degradation of the latency.*

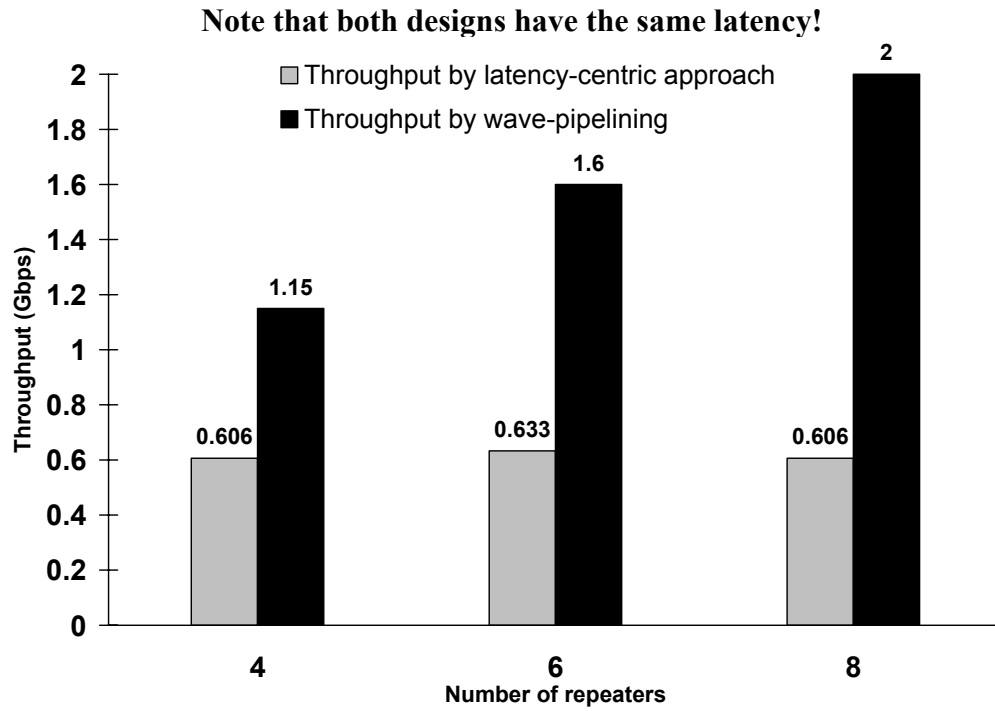


Figure 5.12: Comparison of throughput by latency-centric approach and wave-pipelining.

5.9 Summary

The importance of voltage scaling, repeater insertion, and wire sizing to achieve a high interconnect throughput at the expense of low power, low area, and low via blockage is discussed in this chapter. Using different design metrics, the optimal supply voltage, repeater density, and interconnect dimensions are derived for different applications. The design optimizations shown in this chapter are not based on any assumptions about wiring distributions. As a result, these optimizations are general, simple and they give the designer a lot of flexibility in the design.

The design optimizations show that the optimal supply voltage for low-power interconnect circuits is twice the threshold voltage. It is also shown through the design

optimization of a high-performance cache bus using a throughput-per-energy-area (TPEA) metric that an optimized wave-pipelined interconnect can achieve the necessary throughput at the expense of lower power, power density, and wire area than the LVDS circuit. Because the TPEA optimization is simple and fully scalable, it can be directly applied to any interconnect lengths in a particular technology generation.

The application of wave-pipelining along with latency-centric repeater insertion is proposed in this chapter for the latency-sensitive applications. It is shown that wave-pipelining at the near-optimal latency-centric design can significantly enhance the interconnect throughput without considerably changing the latency, which makes it a very useful technique for such applications.

CHAPTER 6

SYNCHRONIZATION OF DATA ON WAVE-PIPELINED INTERCONNECTS

6.1 Introduction

The preceding chapters have explained the concept of wave-pipelining, discussed the importance of wave-pipelining to enhance interconnect throughput, and presented different methods for interconnect design optimization. It is shown that wave-pipelining is an effective technique to enhance interconnect throughput at the expense of minimal power and area. However, despite all its advantages, one of the key reasons behind the limited use of wave-pipelining in practical applications is the inherent difficulty in data synchronization [16]. Wave-pipelining performs high-speed serialization of the data-bits, which need to be correctly captured at the output. The data on wave-pipelined interconnects are not latched with the clock at regular time intervals, which makes it challenging to synchronize the data with the clock at the output. Therefore, the system-level analysis of wave-pipelining is performed in this chapter to address these timing and interfacing issues.

First, the wave-pipelined interconnect with receiver is compared to the latch-inserted interconnect, which is a conventional pipelining technique for global

interconnects. An analysis of different timing circuits developed in recent years to successfully capture the data on wave-pipelined interconnect circuits is then presented. By modifying these circuits to overcome their limitations, new timing circuits are proposed for the synchronization of data on wave-pipelined interconnects for different system architectures. Finally, a detailed circuit-level analysis of the skew-insensitive retimer circuit for interfacing wave-pipelined interconnects with systems-on-chip (SoC) that use a globally asynchronous locally synchronous (GALS) scheme is performed.

6.2 Comparison of wave-pipelining and latch insertion

Global interconnects in the conventional VLSI systems are pipelined by inserting latches on them. Therefore, in this section, the wave-pipelined interconnect with a simple receiver is compared to the latch-inserted interconnect in the areas of performance, power, and area. The overhead of timing and synchronization circuits is also included in this analysis.

6.2.1 Wave-pipelined interconnect with simple latch receiver

Instead of periodically synchronizing the data to the clock, wave-pipelining suggests latching the data only once at the output of the interconnect. It is assumed for the wave-pipelined interconnect in this section that the clock is sent along with the data. Based on [5], one clock line is sent along with eight data channels, therefore, the area overhead of the clock line per data line is only 12.5%. The clock lines and data lines use the identical interconnect geometry and are expected to experience identical power supply fluctuations and manufacturing and processing variations.

The wave-pipelined interconnect circuit with the latch receiver is shown in Figure 6.1. As seen in Figure 6.1, one data bit is sent in one clock cycle by using simple repeater drivers. The data are synchronized to the clock at the output using a latch, whose construction is also shown in Figure 6.1. Though the clock and the data are expected to reach the output at the same time, any inherent skew between the two is removed at the output.

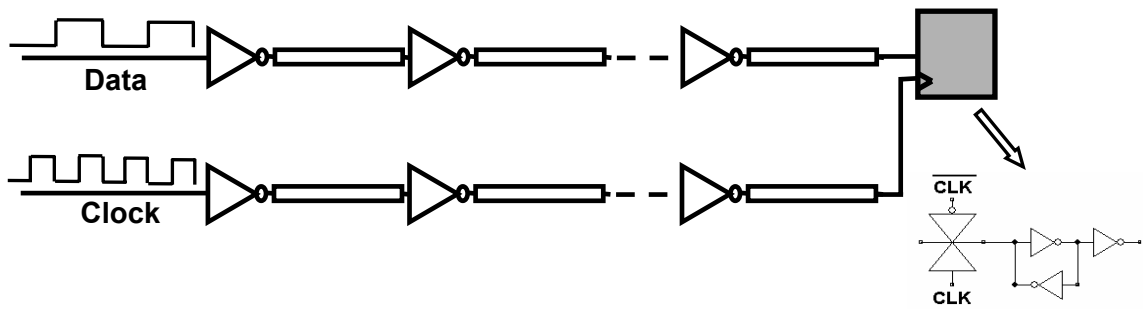


Figure 6.1: Wave-pipelined interconnect with receiver.

It is seen from the clock and data waveforms in Figure 6.1 that when one data bit is sent every clock cycle, the clock line has *twice* the *throughput* compared to that of the data line. Therefore, it is the clock line that operates the maximum throughput in such wave-pipelined interconnect circuits. As a result, the performance on these circuits is mostly limited by the speed at which the clock can be sent.

6.2.2 Latch-inserted interconnect

Like repeaters, latches are periodically inserted on long interconnects, but they are driven by the clock. In one clock cycle, the data are transferred from one latch to the next. A latch-inserted interconnect is shown in Figure 6.2. The same latch that is used as a

receiver for the wave-pipelined interconnect in the previous subsection is used to periodically latch the data, based on its circuit in [71]. This latch circuit is specially designed in [71] to achieve low power.

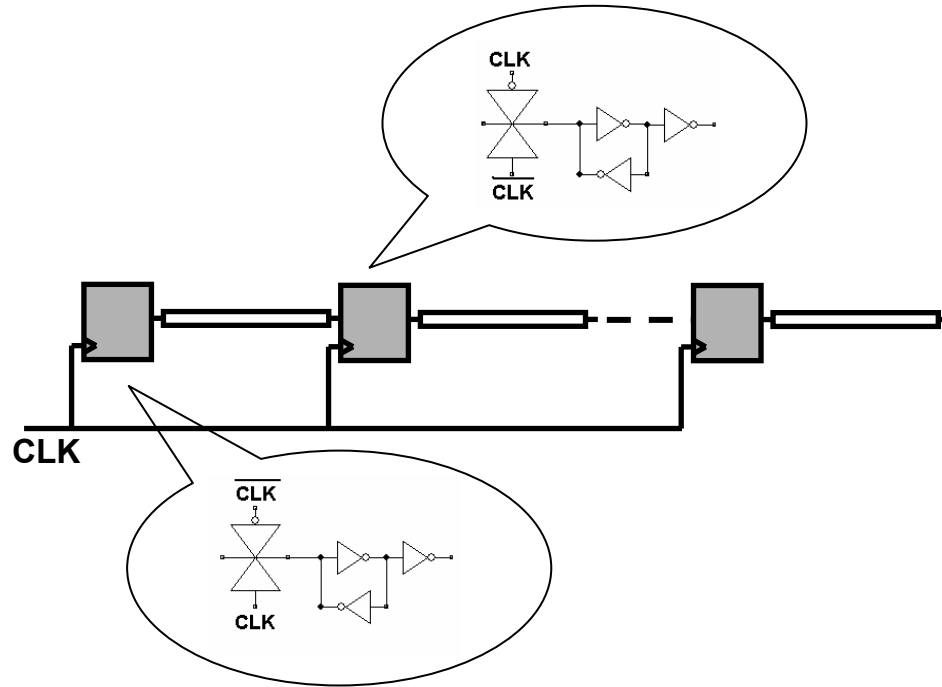


Figure 6.2: Latch-inserted interconnect.

To minimize the latency on latch-inserted interconnects, alternate latches are made sensitive to different levels (HIGH or LOW) of the clock, as seen in Figure 6.2. In the positive half of the clock cycle, the data bit enters the first latch and passes through the transmission gate and inverters onto the interconnect segment. This data bit is captured by the second latch in the negative half of the clock cycle and transmitted on the following interconnect segment.

Though latches can be used to enhance the interconnect performance through pipelining, they consume a lot of power and area [5]. Moreover, the additional set-up and hold constraints associated with the latches not only make it difficult to time the data at every interconnect segment, but also limit the interconnect performance.

6.2.3 Performance of latch-inserted interconnect and wave-pipelined interconnect with receiver

The wave-pipelined interconnect circuit and the latch-inserted interconnect circuit are modeled in HSPICE using a 1 cm long metal-5 interconnect in the 180 nm technology [29], whose dimensions are shown in Table 1.1. A 2 V supply is used for both these circuits. The repeaters and latches have comparable sizes. A repeater scaling factor of 56 is used for the wave-pipelined interconnect circuit, and the latches are sized based on [71]. Results for the performance of these two circuits are shown in Table 6.1.

Table 6.1: Performance comparison of wave-pipelined interconnect with latch-inserted interconnect in 180 nm technology generation.

Number of latches or repeaters	Maximum Throughput (Gbps)		Latency (ns)	
	Latch	WP	Latch	WP
1	1.00	1.11	0.50	0.42
5	2.38	1.92	1.05	0.71
10	2.86	2.44	1.75	0.86
20	3.00	3.13	3.33	1.04

Latch: latch-inserted interconnect and WP: wave-pipelined interconnect.

It is seen in Table 6.1 that the wave-pipelined interconnect results in a comparable performance as the latch-inserted interconnect, but it results in a considerably lower latency. For 20 pipeline stages, the latency of the wave-pipelined interconnect is less than one third of that of the latch-inserted interconnect, without any loss of throughput performance.

6.2.4 Comparison of latch-inserted interconnect with wave-pipelined interconnect under a constant throughput constraint

It is seen from Table 6.1 that for the given transistor and interconnect dimensions, the saturation throughput for the latch-inserted interconnect is almost 3 Gbps. Therefore, the latch-inserted interconnect circuit is compared to the wave-pipelined interconnect circuit for a constant throughput constraint of 3 Gbps. As seen in Table 6.1, a 3 Gbps throughput is achieved on the 1 cm long metal-5 interconnect [29] in the 180 nm technology by inserting 20 latches. This throughput can be achieved on the wave-pipelined interconnect by inserting 18 repeaters. The schematic diagrams for these two circuits are shown in Figure 6.3, and these circuits are compared in Table 6.2 in terms of power and area for a constant throughput performance of 3 Gbps. The results in Table 6.2 are based on HSPICE simulations.

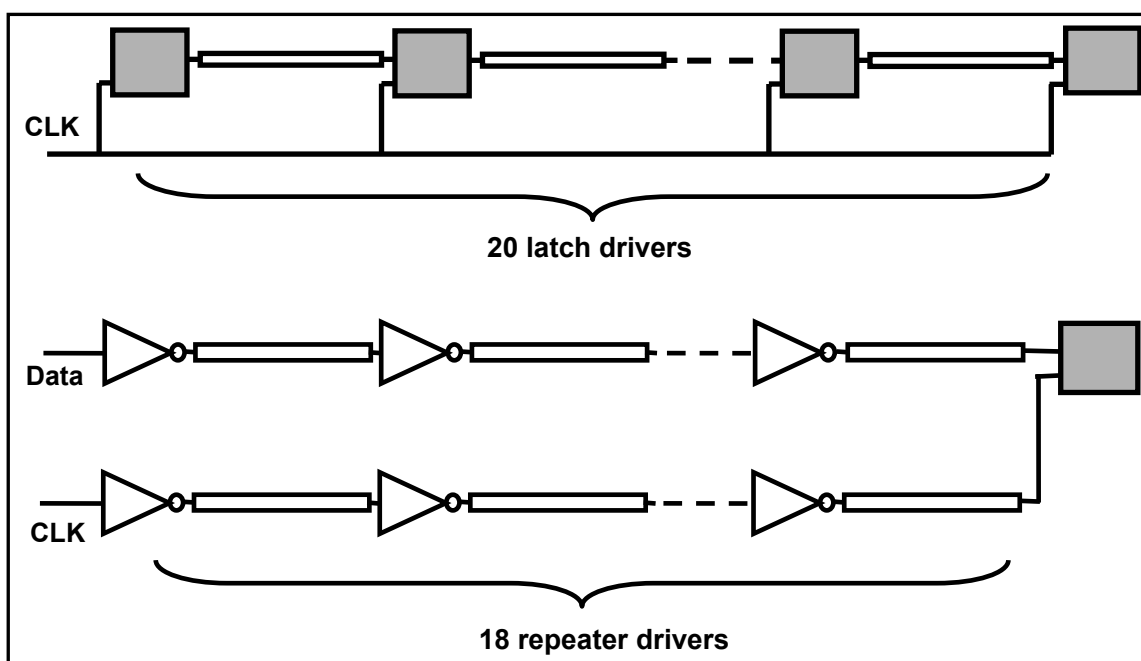


Figure 6.3: Schematic representation of latch-inserted interconnect and wave-pipelined interconnect with receiver.

Table 6.2: Comparison of latch insertion and wave-pipelining for a constant throughput of 3 Gbps.

Quantity	Latch insertion	Wave-pipelining
Number of latches/repeaters	20	18
Throughput	3 Gbps	
Latency	3.33 ns	0.97 ns
Total power	6.23 mW	3.77 mW
Wire area	1.6E-04 cm ²	1.8E-04 cm ²
Silicon area	5.07E-05 cm ²	1.05E-05 cm ²

It is seen in Table 6.2 that for a constant throughput performance of 3 Gbps, wave-pipelining results in a 70% reduction in latency, a 40% reduction in total power, and an 80% reduction in silicon area compared to latch insertion. The total power in Table 6.2 includes all of the dynamic, leakage, and short-circuit power, and assumes an activity factor of 0.1 [28]. It is observed that the leakage power for the latch-inserted interconnect circuit (86 μ W) is almost four times higher compared to that for the wave-pipelined interconnect circuit (22 μ W). The power and area for the wave-pipelined interconnect in Table 6.2 also include the overhead for the clock line. The values of power and area are calculated based on the assumption that one clock line is sent along with eight data channels.

The comparison in Table 6.2 does not assume any voltage scaling or wire sizing for the wave-pipelined interconnect. Even if a smaller supply voltage or reduced wire sizes are used for the wave-pipelined interconnect, the same throughput could be obtained by inserting more repeaters. Such a design can further reduce power and also reduce wire area, thereby compensating for the overhead of the clock line. Thus, the larger design space provided by wave-pipelining not only makes the wave-pipelined interconnect design more flexible, but it also gives the designer a better control over power, area, and performance of the circuit.

6.3 Existing solutions for synchronizing data on wave-pipelined interconnects

As shown in the previous section, the data on the wave-pipelined interconnect can be latched at the output using the clock, which is sent along with the data. As also seen in the previous section, this is a very simplistic timing mechanism and results in less area and power overhead than the traditional latch-inserted interconnect. However, the wire area overhead of the clock line and the underutilization of the data line (because the data line operates at *half* its maximum capacity) are two drawbacks of this scheme. Either or both of these drawbacks are attempted to be removed in the synchronization circuits proposed by some researchers to capture the data on wave-pipelined interconnects [5], [16]. One approach suggests sending the clock along with the data (where one data bit is transmitted in every *half* clock cycle) and latching the data using a flip-flop-based receiver [5], whereas the other approach suggests locally generating the clock and latching the data using a phase-locked loop (PLL) [16]. The mechanisms behind these two approaches, their advantages, and shortcomings are discussed in this section.

6.3.1 Sending clock along with data and latching data using flip-flops

Xu and Wolf have suggested sending the clock along with data to latch the output data on wave-pipelined interconnects using a flip-flop-based receiver [5]. The receiver circuit proposed in [5] is shown in Figure 6.4. In the 250 nm technology that is used for validation in [5], one clock line is sent along with every eight data channels, and one data bit is sent in every half cycle of the clock.

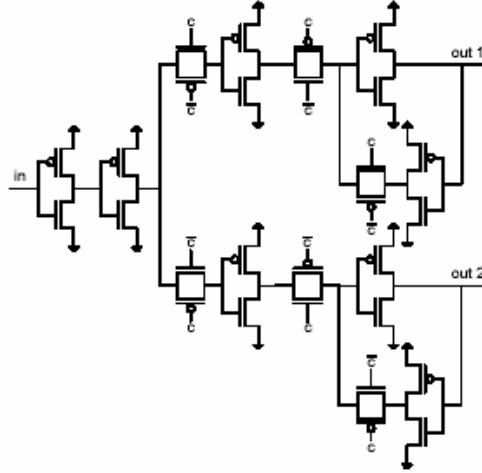


Figure 6.4: Flip-flop-based receiver in [5].

In Figure 6.4, the node ‘in’ denotes the output of the interconnect, which acts as the input to the receiver, and the nodes ‘out1’ and ‘out2’ denote the outputs synchronized to the two halves of the clock cycle. The signals ‘c’ and ‘ \bar{c} ’ denote the clock and its inverse, respectively. Because the clock is sent along with the data, it is expected that they will experience identical manufacturing and power supply fluctuations and reach the output at the same time.

The data bit is pumped into the upper flip-flop in Figure 6.4 in the positive half cycle of the clock. In the negative half cycle of the clock, this data bit is transferred to the output, and the following data bit is pumped into the lower flip-flop. Therefore, after a delay of a half clock period, the data bits are available at the output nodes. The feedback loops at the output prevent the data from dissipating [5].

The primary advantage of this circuit is that the data interconnects can operate at the maximum possible speed because the clock and the data signals have identical pulsewidths. This is called ‘double data rate’ (DDR) mechanism. In this chapter, the

logic circuits that process one data bit every half cycle are also referred to as DDR cores. Unlike this mechanism, if the drivers and receivers send/receive one data bit every clock cycle (as in most conventional systems), it is called 'single data rate' (SDR) mechanism, and such logic circuits are referred to as SDR cores in this work.

The DDR approach is used in some memory controllers (e.g., memory controller of Freescale Semiconductor's PowerQUICC III processor) and certain on-chip buses in high-performance microprocessors such as Itanium [72] and is being researched upon as a high-performance solution for future processor systems. However, most cores in the processor typically use an SDR approach. To use the receiver circuit of Figure 6.4 to connect SDR cores, the data from two input SDR cores can be multiplexed on a single DDR interconnect. The flip-flop-based receiver circuit in Figure 6.4 can split the data and feed them to their respective SDR receiver cores. This idea is further developed in Section 6.4.2.

The HSPICE simulation of a 180 nm metal-5 [29] wave-pipelined interconnect using the receiver in Figure 6.4 shows that it can sustain a 4 Gbps throughput on the interconnect with 10 repeaters per cm. It is observed that the throughput is primarily limited by the response times of the flip-flops. The authors of [5] also hint at the use of buffers for a possible reading of the output data by another clock in a different clock domain. However, this idea is not further developed by any logic or circuit design. Therefore, the applicability of the receiver in [5] is restricted to the fully synchronous systems.

6.3.2 Locally generating clock and latching data using PLL

To avoid the area overhead of the clock line, Zhang et al. suggest locally generating the clock at the receiver using a PLL [16]. PLL-based receivers have been successfully implemented in the off-chip communication, and [16] attempts to use the same technique for the on-chip wave-pipelined interconnects. In this synchronization scheme, the phase of the locally generated receiver clock is aligned with the incoming data using a PLL. A 2.8 cm long 180 nm wave-pipelined global interconnect is reported to achieve a 2 Gbps throughput in [16]. Figure 6.5 shows the PLL-based receiver in [16].

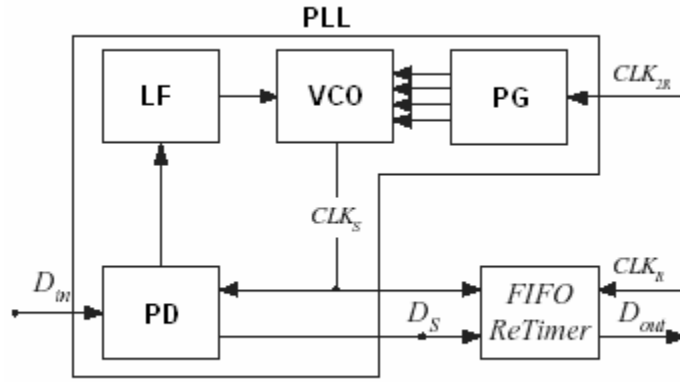


Figure 6.5: PLL-based receiver in [16].

In Figure 6.5, the term LF stands for a loop filter, VCO stands for a voltage-controlled oscillator, PG stands for a phase generator, and PD stands for a phase detector. VCO is a digital counter assisted by a 4:1 multiplexer, which selects the input clock that has a minimal phase difference with the incoming data. The phase detector and the loop filter then align the data and the clock with each other. However, this process can take

multiple clock cycles, which not only increases the latency, but also necessitates the use of the header bits at the beginning of data.

The area and power analysis of the wave-pipelined interconnect with PLL-based receiver is also presented in [16]. Because of the enormous overhead of the PLL-based receiver, the 180 nm wave-pipelined circuits are shown to be area- and power-efficient than the conventional latch-inserted interconnects only when the interconnect length is greater than a critical length of 0.65 cm. Moreover, because the phase detector, loop filter, and other analog devices are used in the digital system, there could be some circuit-level interfacing problems, which are not fully explored in [16]. The use of a first-in-first-out (FIFO) retimer circuit is also suggested in [16] for a possible reading of the data by another clock in a different clock domain. The logic design of the retimer circuit is discussed in [16], and the retimer circuit is synthesized using standard cells.

6.4 Simplified receiver for fully synchronous systems

To overcome the shortcomings of the existing receiver circuits, a simple new receiver is proposed in this research to synchronize data to the clock on the wave-pipelined interconnect. Similar to [5], one clock line is sent along with eight data channels and instead of carrying one data bit every clock cycle, the interconnect carries it every *half* clock cycle. Therefore, both the clock and the data lines can be operated at the maximum possible throughput, and the data interconnect always operates in the DDR mode. However, though the interconnect carries one data bit every half clock cycle, the drivers/receivers may produce/consume it every half or one clock cycle. Therefore, the synchronization scheme is considered for two scenarios as follows:

1. The *DDR* cores are connected with wave-pipelined DDR interconnects.
2. The *SDR* cores are connected with wave-pipelined DDR interconnects.

6.4.1 Interfacing DDR cores with wave-pipelined DDR interconnects

Leaving the basic principle behind the receiver in Figure 6.4 [5] unchanged, the receiver circuit is modified as shown in Figure 6.6, to obtain a higher throughput and better signal integrity. Figure 6.6 shows the receiver for a scenario where the drivers and receivers operate with the DDR mechanism, i.e., they send/receive one data bit every half clock cycle. This circuit is further simplified in the following subsection to interface with drivers and receivers operating with the SDR mechanism.

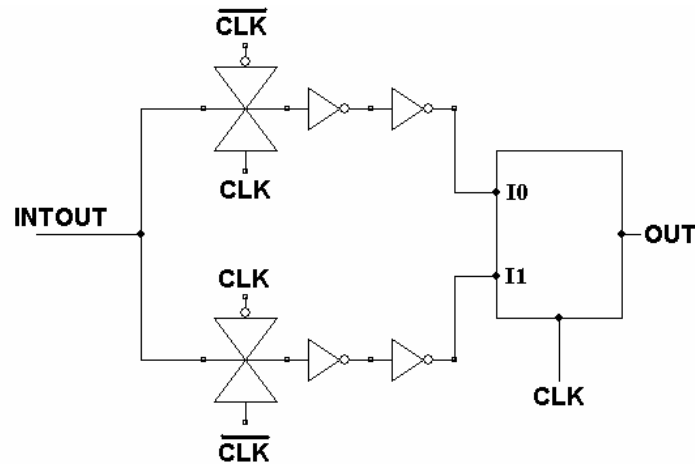


Figure 6.6: Receiver for wave-pipelined interconnect connecting DDR cores.

As seen in Figure 6.6, simple transmission gates are used to capture the data at the output. By avoiding the feedback loop, the simple receiver circuit in Figure 6.6 achieves a high performance and maintains good signal integrity. This receiver circuit is used to

perfectly align the output data with the clock, thereby removing any skew between the two that is inherent or generated during the transmission on the interconnect.

The interconnect is driven by a multistage inverter driver. The clock and data are sent on their respective interconnects at the same time. The node 'INTOUT' in Figure 6.6 denotes the output of the interconnect, which acts as the input to the receiver circuit. The terms 'CLK' and ' $\overline{\text{CLK}}$ ' denote the clock and its inverse, respectively. As seen in Figure 6.6, the output of the interconnect is captured by the upper transmission gate and stored into I0 in the positive half of the clock cycle. Similarly, the data bit at INTOUT is captured by the lower transmission gate and stored into I1 in the negative half of the clock cycle. I0 and I1 are the inputs of the multiplexer, and the clock signal acts as its control signal.

The first half of the clock cycle is positive, when the data bit at I1 (that is not yet valid) gets transferred to the node 'OUT'. In the negative half cycle, the valid data bit at I0 is transferred to the output, whereas in the following positive half cycle, the valid data bit at I1 is transferred to the output. Thus, the data bits get perfectly synchronized with the clock by the receiver circuit after the delay of a half clock period. It should be noted that writing into the multiplexer and reading from the multiplexer occurs in *different* halves of the clock, which not only provides the necessary isolation, but also enhances the speed at which the receiver circuit can operate.

The simple expression for the multiplexer output is given as

$$\text{OUT} = \text{I0} \cdot \overline{\text{CLK}} + \text{I1} \cdot \text{CLK} . \quad (6.1)$$

The gate-level simplification of (6.1) results in

$$\text{OUT} = (\overline{\text{I0}} \text{ NOR } \text{CLK}) \text{ OR } (\overline{\text{I1}} \text{ NOR } \overline{\text{CLK}}) . \quad (6.2)$$

Based on (6.2), the circuit details for the multiplexer are provided in Figure 6.7. The aspect ratios for all the transistors are shown next to them.

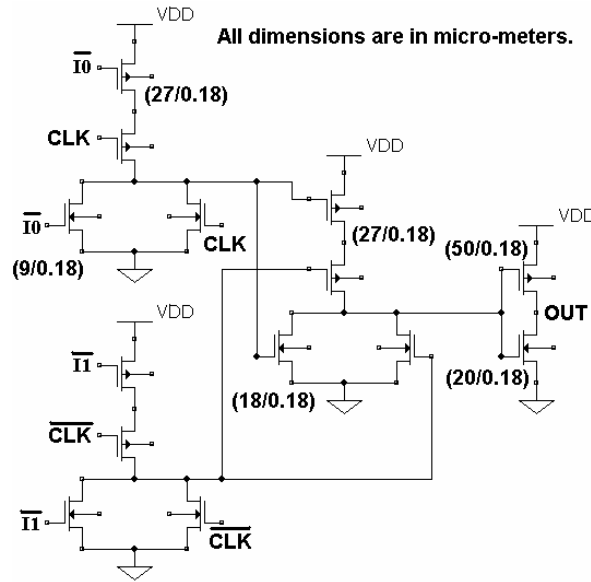


Figure 6.7: Circuit diagram for 2:1 multiplexer in 180 nm technology.

HSPICE simulations show that a 180 nm metal-5 interconnect [29], whose dimensions are shown in Table 1.1, with the receiver in Figure 6.6 can sustain a 4.54 Gbps throughput with 10 repeaters. This throughput is almost 15% higher and the signal integrity is better than that obtained by using the receiver in Figure 6.4 for the same interconnect circuit. Because the clock is sent along with the data using an identical interconnect geometry, it is expected to reach the output at the same time as data. Nevertheless, HSPICE simulations show that this circuit can tolerate up to 10-12% skew between the *clock* and the *data*.

The timing diagram for this scenario is shown in Figure 6.8 corresponding to a 2.27 GHz clock frequency. As seen in Figure 6.8, the receiver effectively aligns the data to the clock, thereby removing the skew between them. It is reemphasized by the waveforms in Figure 6.8 that *a 2.27 GHz clock frequency translates into a 4.54 Gbps clock throughput*. When the clock is routed using an identical geometry as the data lines, both the clock line and the DDR interconnect nominally operate at the maximum channel throughput, i.e., 4.54 Gbps in the present case. Because the driver-receiver cores also operate with the DDR mechanism, the data bits are also generated at the rate of 4.54 Gbps. Thus, this receiver circuit supports high-speed communication on the wave-pipelined interconnects and is suitable to be integrated into any synchronous system with DDR capability.

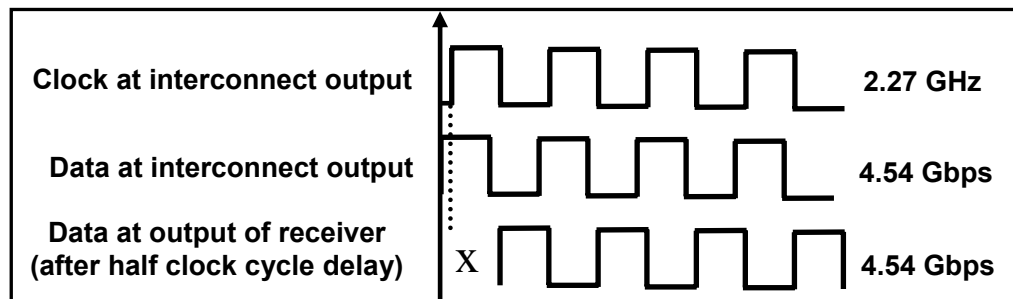


Figure 6.8: Timing waveform for DDR cores interfaced with DDR interconnect.

6.4.2 Interfacing SDR cores with wave-pipelined DDR interconnects

The receiver circuit in Figure 6.6 assumes that the drivers and the receivers use the DDR mechanism. However, the simplified version of this circuit can be used when the cores connected by the wave-pipelined interconnects are SDR cores, i.e., the drivers and receivers send/receive the data every clock cycle.

By using a 2:1 multiplexer at the driver side, the data from two adjacent drivers can be alternately sampled. The two drivers would generate one data bit with every new clock cycle, but the data would be multiplexed such that one data bit appears on the interconnect after every *half* clock cycle. The wave-pipelined interconnect can then carry the data in a DDR fashion as before, which can be split at the receiver by using the positive and negative levels of the clock and fed to the respective receivers. The schematic diagram for this scenario is shown in Figure 6.9.

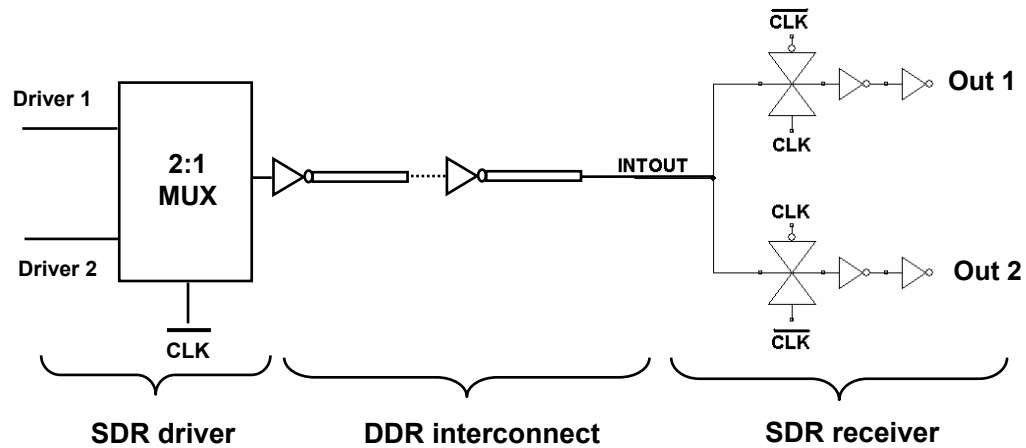


Figure 6.9: Interfacing SDR cores with wave-pipelined DDR interconnect.

As seen in Figure 6.9, because of the multiplexing, two interconnects have been replaced by a single interconnect that operates in a DDR fashion. Therefore, there is a 50% reduction in the wire area for the data lines. Though there is an overhead of the clock line, one clock line is sufficient for eight data channels, and the wire area reduction resulting from multiplexing is significantly greater than the clock line overhead. Therefore, this scheme fits well in the scenarios where the cores operate with the conventional SDR mechanism and the system is primarily wire-limited. A wave-

pipelined multiplexing (WPM) technique that uses a similar approach, but does not send the clock along with the data, is presented in [28].

The timing diagram for the present scenario is shown in Figure 6.10 corresponding to a 2.27 GHz clock frequency. It is seen in Figure 6.10 that though the data bits are generated at the rate of 2.27 Gbps, the interconnect carries them at 4.54 Gbps. The multiplexing technique used in this scenario not only allows the data lines to operate at the maximum possible throughput, but also results in a 50% reduction in the wire area.

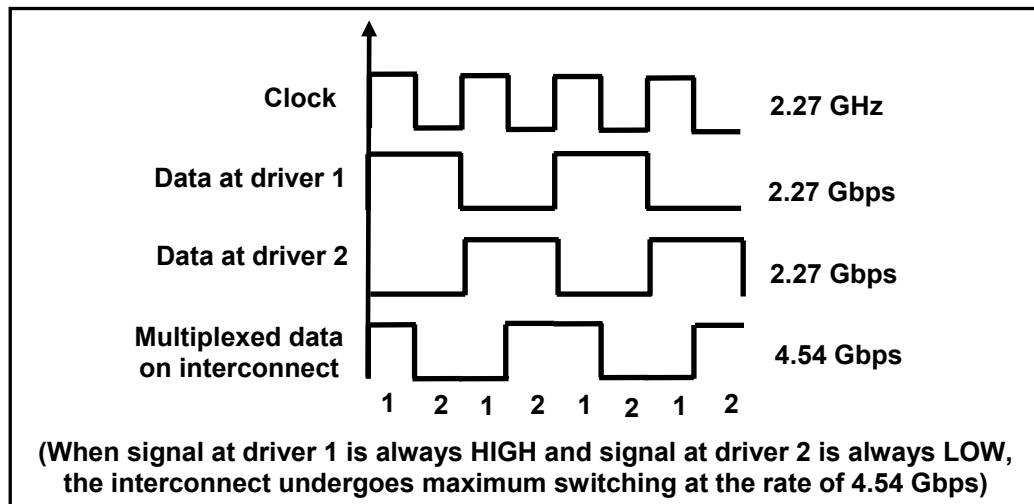


Figure 6.10: Timing waveform for SDR cores interfaced with DDR interconnect.

6.5 Globally asynchronous locally synchronous (GALS) systems

With the integration of multiple cores onto a single chip, the VLSI designs are exhibiting a possible trend toward a system-on-chip (SoC). The SoC circuits are expected to work at very high frequencies and have more than one billion transistors on a single chip [2]. Moreover, the area and power requirements are more stringent for the SoC

circuits. Therefore, it is necessary to achieve the required high performance at the expense of minimal power and area. Wave-pipelining is shown to meet these requirements in the analyses in Chapter 5. Therefore, SoC interconnects are a great opportunity to apply wave-pipelining.

In the large SoC circuits, the data and control signals are expected to take multiple clock cycles to travel between the farthest cores [73]. To support high-speed communication between multiple cores, a globally asynchronous locally synchronous (GALS) scheme could be used [74], [75]. In the GALS scheme, global interconnects need to communicate between different clock domains. Therefore, the receiver of a global interconnect may operate with a clock having a different phase (or possibly, a different frequency) than that at the driver side. In this research, the focus is on cores that have the same clock frequency, but a completely random phase, i.e., a random skew between the source clock and the locally generated destination clock.

To use wave-pipelining on the SoC global interconnects in the GALS scheme, it is important to design a receiver that can latch the incoming data with a clock that has a completely random phase w.r.t. the source clock. To interface wave-pipelining with GALS, both [5] and [16] suggest a possible use of a FIFO retimer circuit, in which the data is written into by the source clock and read out by the receiver clock. Though [16] has synthesized such FIFO retimers using standard gate libraries, a careful literature review shows that the full-custom circuit design or circuit validation with distributed RLC interconnect models has not been published in the past for the retimer circuit. Therefore, this section presents a skew-insensitive retimer circuit to latch the data on the wave-pipelined interconnects in the GALS scenario.

6.5.1 Skew-insensitive retimer circuit

As in Section 6.4, the proposed scheme recommends sending the source clock along with data. Because the clock and the data use an identical interconnect geometry, they are expected to experience the same manufacturing and process variations and power supply fluctuations and reach the output at almost the same time. To analyze the maximum possible performance that can be obtained in the GALS scenario, the driver and receiver cores are assumed to operate with the *DDR* mechanism, and the circuit schematic for this implementation is shown in Figure 6.11. However, the retimer circuit for SDR cores is also discussed later in Section 6.5.5, which includes a discussion of the modification to the timing and control circuits.

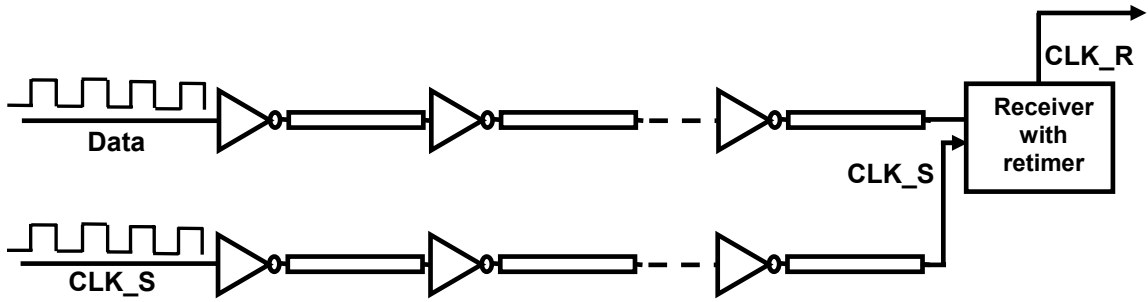


Figure 6.11: Circuit schematic for DDR interconnects with DDR retimer for GALS systems.

In the proposed receiver circuit, the data bits are first captured with the help of the source clock, CLK_S . The retimer circuit then aligns the data to a local receiver clock, CLK_R , which could be significantly skewed w.r.t. CLK_S . Figure 6.12 presents the circuit-level implementation details for the retimer circuit. The aspect ratios of all the MOSFETS are shown in parentheses in Figure 6.12.

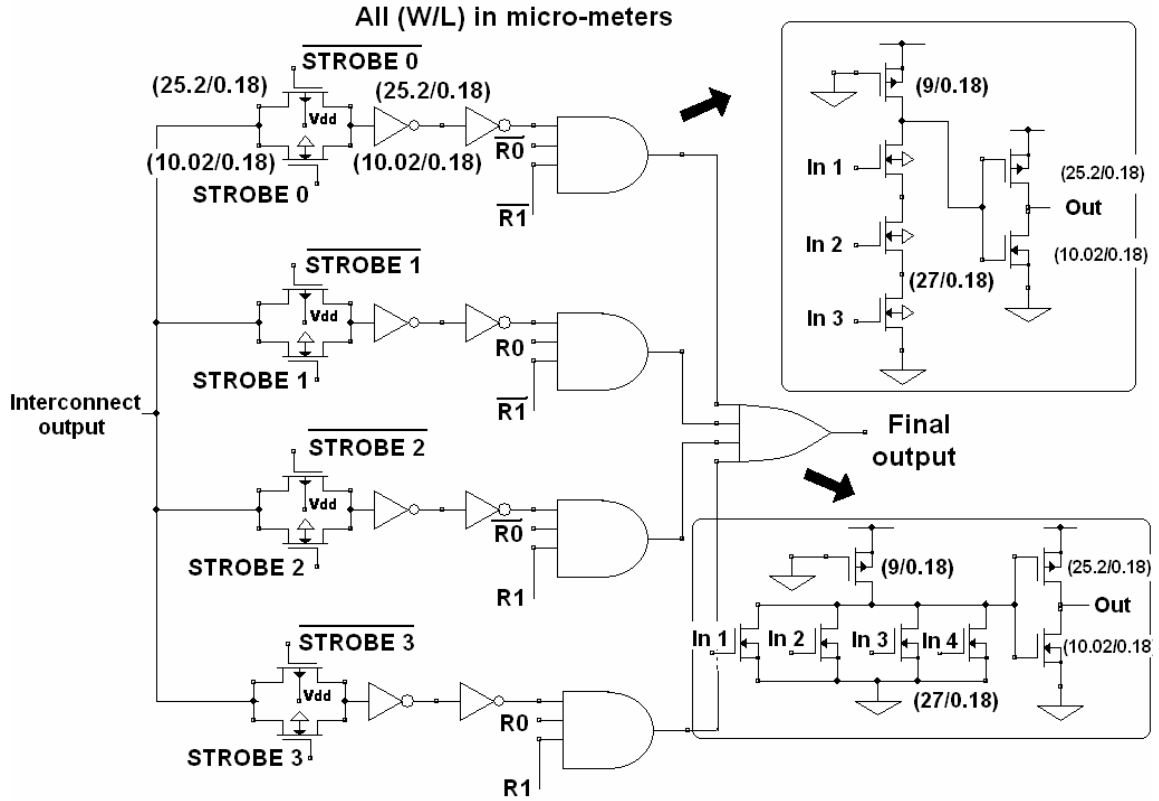


Figure 6.12: Retimer circuit for wave-pipelined interconnects in GALS scenario.

Upon its arrival, CLK_S fires a synchronous counter at the receiver into a periodic generation of the signals S_1S_0 , which assume the values of 00, 01, 10, 11, one with every new *half* cycle of CLK_S. The signals S_1S_0 then generate the necessary STROBE signals, and the simple circuits to generate these STROBE signals are shown in Section 6.5.2. CLK_S also turns on another synchronous counter at the output, which then generates the signals R_1R_0 with every *half* cycle of the locally generated CLK_R and controls the 4:1 multiplexer. This counter generates R_1R_0 in the order 11, 00, 01, 10, thereby avoiding the simultaneous writing and reading at the same input of the multiplexer. It is this counter that aligns data with the local receiver clock.

The multiplexer is implemented by the combination of AND-OR gates, as shown in Figure 6.12. The multiplexing action at the receiver starts with the *first positive half of CLK_R after receiving CLK_S*, and the output of the multiplexer is always valid beginning the *second* half of CLK_R, when R₁R₀ become 00.

Writing into the multiplexer (controlled by S₁S₀) and reading from the multiplexer (using R₁R₀) are thus interleaved by a half clock cycle. To further explain this interleaving, Table 6.3 shows the values of control signals for the case where there is no skew between CLK_S and CLK_R. In Table 6.3, M0, M1, M2, and M3 denote the four inputs of the multiplexer.

Table 6.3: Values of control signals when CLK_R is in phase with CLK_S.

S ₁	S ₀	STROBE0	STROBE1	STROBE2	STROBE3	Data written into	R ₁	R ₀	Data read from
0	0	1	0	0	0	M0	1	1	M3 (x)
0	1	0	1	0	0	M1	0	0	M0
1	0	0	0	1	0	M2	0	1	M1
1	1	0	0	0	1	M3	1	0	M2

x: don't care condition (data not valid)

Because of this interleaving, the retimer circuit tolerates a maximum of 360° skew between CLK_S and CLK_R, which is the maximum possible skew that can exist between any two clocks. The retimer circuit thus facilitates latching of the data by an arbitrary receiver clock with any phase difference w.r.t. the source clock. The values of R₁R₀ corresponding to different phase differences between CLK_S and CLK_R are shown in Table 6.4. Regardless of the phase difference, the data are always valid beginning the second half of CLK_R after the synchronous counter starts generating R₁R₀.

Table 6.4: Design for skew tolerance of 0° to 360° between CLK_S and CLK_R.

Writing of data with CLK_S	Reading of data with CLK_R		
	Phase difference 0°	Phase difference 180°	Phase difference 360°
M0	M3	X	X
M1	M0	M3	X
M2	M1	M0	M3
M3	M2	M1	M0

X: don't care condition (data not valid)

6.5.2 Generation of control signals for retimer circuit

The accuracy of the retimer circuit depends on the timing of control signals S_1S_0 and R_1R_0 . Therefore, the construction of the synchronous counters that generate these signals is discussed in detail in this subsection. One synchronous counter generates S_1S_0 in the order 00, 01, 10, 11, and the other synchronous counter generates R_1R_0 in the order 11, 00, 01, 10. For these sequences, the required waveforms of S_1S_0 and R_1R_0 are shown in Figure 6.13. The simple circuits to generate STROBE signals from S_1S_0 are shown in Figure 6.14.

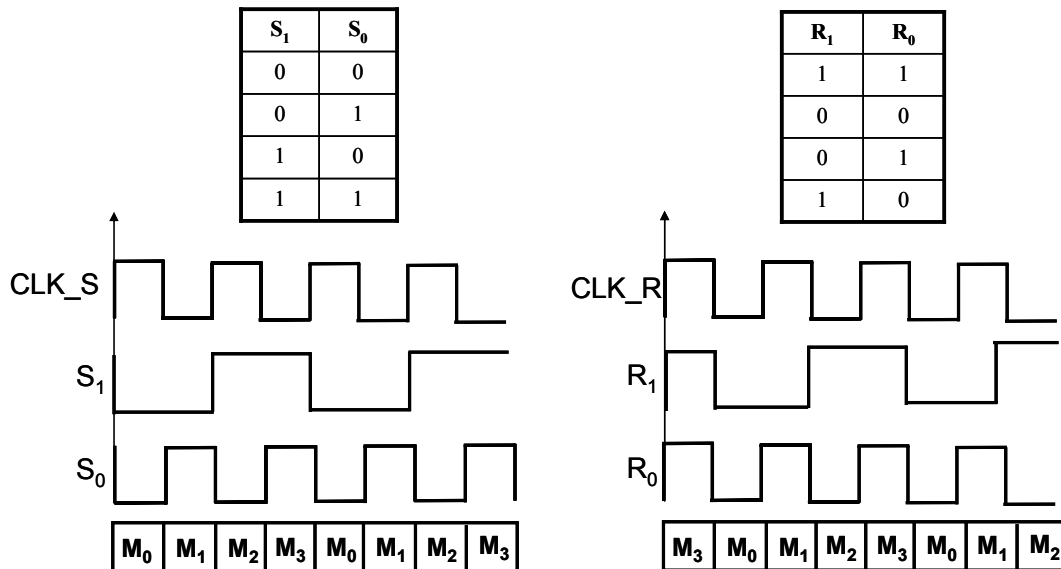


Figure 6.13: Waveforms for S_1S_0 and R_1R_0 .

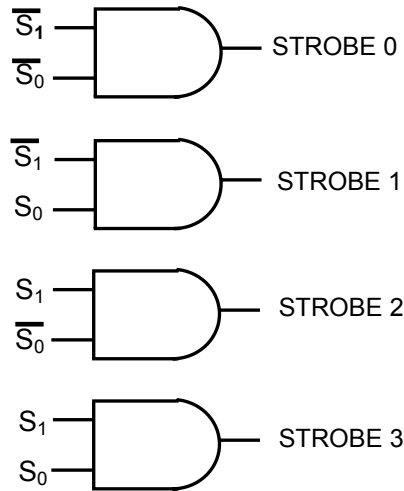


Figure 6.14: Simple AND gates used for generating STROBE signals.

An observation of Figure 6.13 shows that the signals S_0 and R_0 can be directly generated from their respective clock signals. However, the signals S_1 and R_1 have half the frequency compared to the clock signals. Therefore, a divide-by-2 latch, which is better known as a toggle latch, is used to obtain S_1 and R_1 from the clock signals. The construction of the toggle latch is shown in Figure 6.15. The transistor aspect ratios are identical for both latches shown in Figure 6.15. The aspect ratios of transistors in the AND gates and inverters in Figure 6.15 are similar to those in Figure 6.12. The two NOR gates in Figure 6.15 use an NMOS width of $9\text{ }\mu\text{m}$, but they use different PMOS widths of $12\text{ }\mu\text{m}$ and $27\text{ }\mu\text{m}$ to avoid the metastable condition.

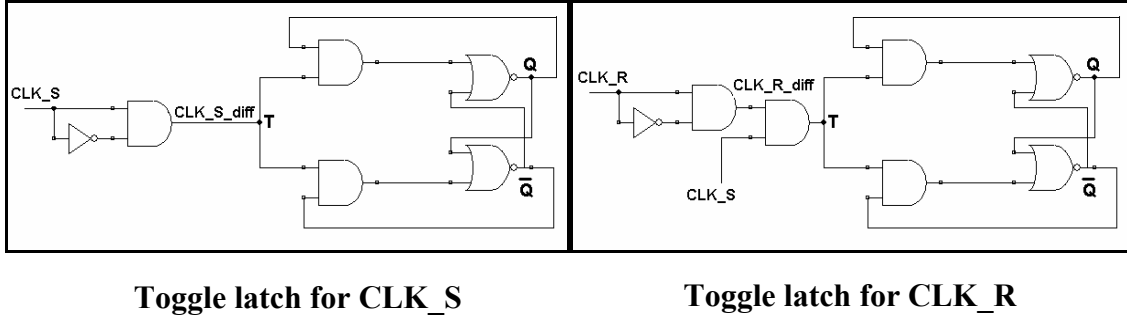


Figure 6.15: Toggle latches for generation of control signals.

As seen in Figure 6.15, the clock signals are logically multiplied with their slightly delayed complements to generate pulses corresponding to the *positive* edges of the clock. These pulses are denoted as CLK_S_diff and CLK_R_diff in Figure 6.15. As seen in Figure 6.15, CLK_S_diff directly acts as the T input for the first toggle latch, whereas CLK_R_diff is logically multiplied with CLK_S to generate the T input for the second toggle latch. The Q and \bar{Q} outputs have a similar waveform as the input clock, but the frequency is halved. The signals S_1 and R_1 are obtained from Q and \bar{Q} outputs of the counter.

Both Q and \bar{Q} get delayed w.r.t. the input clocks because the toggle latches need some time to process the inputs to generate these signals. Therefore, the signals S_0 and R_0 , which are directly obtained from the clock, and the data are deliberately delayed to match with S_1 and R_1 , for the correct operation of the circuit. It is important to note that the incoming data bit needs to be stable until the necessary S_1S_0 and the corresponding STROBE signals are generated to sample it. Therefore, PW_{min} for the data could be limited by the control mechanism rather than the interconnect circuit itself.

6.5.3 HSPICE simulation of wave-pipelined interconnect with retimer circuit

A 1 cm long metal-5 interconnect [29] with 10 repeaters, which uses the retimer circuit shown in Figure 6.12, is simulated in HSPICE using 180 nm level-49 MOSIS transistor models [48]. HSPICE simulations are performed for several different input waveforms and the circuit is successfully tested for up to a 360° skew between CLK_S and CLK_R. The simulated interconnect circuit achieves a maximum throughput of 4 Gbps. If dynamic delay effects are experienced on data lines and not on clock lines, this circuit also tolerates a 20% skew between the *input data* and the *source clock*.

To better explain various timing considerations, Figure 6.16 shows waveforms for important data and control signals in the retimer circuit corresponding to a random data input. It is seen in Figure 6.16 that S_1 and S_0 are generated after the arrival of CLK_S at the output, which then generate the necessary STROBE signals, one in every half cycle of CLK_S. R_1R_0 are generated in synchronization with the first positive transition of CLK_R after the arrival of CLK_S. As seen in Figure 6.16, the output data bits are valid beginning the second half cycle of CLK_R after CLK_S arrives at the output.

As discussed in the previous subsection, generating control signals for the multiplexer in the retimer circuit limits the throughput performance of this interconnect circuit to 4 Gbps. However, a 4 Gbps throughput is still significantly higher than that obtained by the optimal latency-centric repeater insertion (1.32 Gbps) on the same interconnect. Therefore, the retimer circuit discussed in this section provides a fast and robust scheme for the synchronization of data on wave-pipelined interconnects in the GALS scenario.

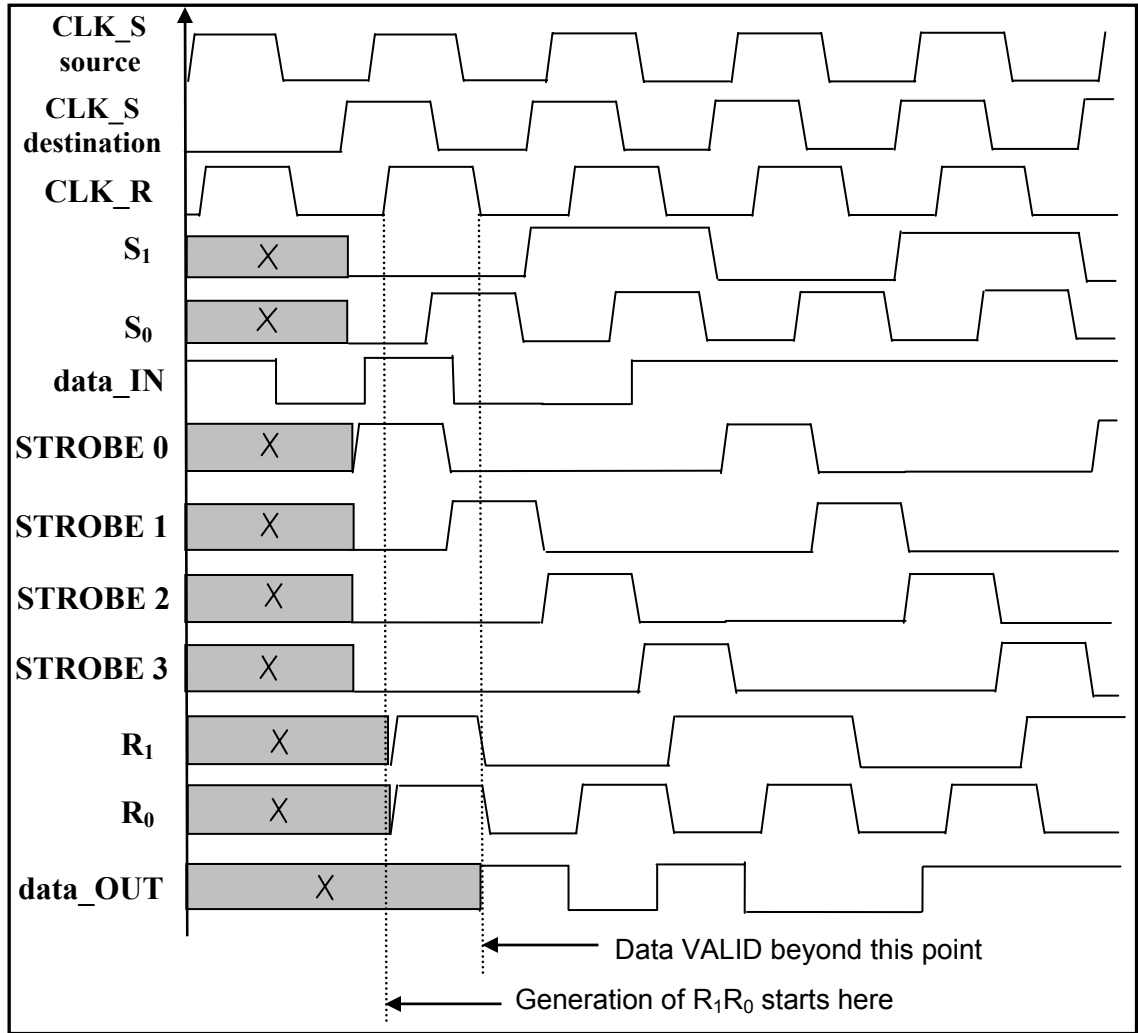


Figure 6.16: Waveforms for various data and control signals in retimer circuit.

6.5.4 Area and power for wave-pipelined interconnect circuit

It is seen in the previous subsection that the 1 cm long 180 nm metal-5 interconnect achieves a 4 Gbps throughput with 10 repeaters. With the assumption that one clock line is sufficient for every eight data channels [5], this subsection analyzes the power and area overhead of the retimer circuit for a throughput performance of 4 Gbps. Table 6.5 provides various design details and values of important parameters for this wave-pipelined interconnect circuit.

Table 6.5: Design details for wave-pipelined interconnect circuit

Interconnect length	1 cm
Interconnect dimensions	$w = 0.8 \mu\text{m}$, $h = 1.6 \mu\text{m}$, $s = 0.8 \mu\text{m}$, $t = 1.6 \mu\text{m}$
Number and size of repeaters	10 repeaters, $W_n = 10.02 \mu\text{m}$ and $W_p = 25.2 \mu\text{m}$
Throughput	4 Gbps
Latency at interconnect output	0.76 ns
Total power	19.7 mW
Wire area	$1.8\text{E-}04 \text{ cm}^2$
Silicon area	$5.6\text{E-}05 \text{ cm}^2$ (Repeaters: $0.88\text{E-}05 \text{ cm}^2$, Toggle latches: $1.72\text{E-}05 \text{ cm}^2$, Delay/STROBE circuitry: $2.20\text{E-}05 \text{ cm}^2$, Mux: $0.80\text{E-}05 \text{ cm}^2$)
Clock skew tolerance	360°

The total power shown in Table 6.5 corresponds to an activity factor of 0.1 [28] and includes all of the dynamic power, leakage power, and short-circuit power. The power and area shown in Table 6.5 include the overhead resulting from the clock signal, synchronous counters, delay circuitry in the retimer circuit, etc.

Though the power and silicon area requirements are relatively high, it is at the cost of this power and area overhead that the interconnect circuit becomes compatible with the GALS system and tolerates any skew between the source clock and the destination clock. Moreover, these results are specific to the DDR cores connected with DDR interconnects, but if the cores operate with an SDR mechanism, the multiplexing suggested in Section 6.5.4 can reduce the wire area and repeater area by 50%. To further optimize power and area for a given performance, optimal voltage scaling, repeater insertion, and wire sizing can also be performed as discussed in Chapter 5.

6.5.5 Retimer circuit for SDR driver and receiver cores

If the driver and receiver cores operate with an SDR mechanism (i.e., they produce or consume one data bit every half clock cycle), the synchronous counters,

which generate the control signals S_1S_0 and R_1R_0 , need to be slightly modified. Regardless of whether the data bits are received by a simple latch as shown in Figure 6.1 (i.e., SDR interconnects, no multiplexing) or they are multiplexed at the driver side and split at the receiver side as shown in Figure 6.9 (i.e., DDR interconnects), the same retimer circuit can be used at the receiver side. *The control signals for such a mechanism need to be generated every clock cycle, and not every half clock cycle.* The waveforms for the clock and the control signals are shown in Figure 6.17.

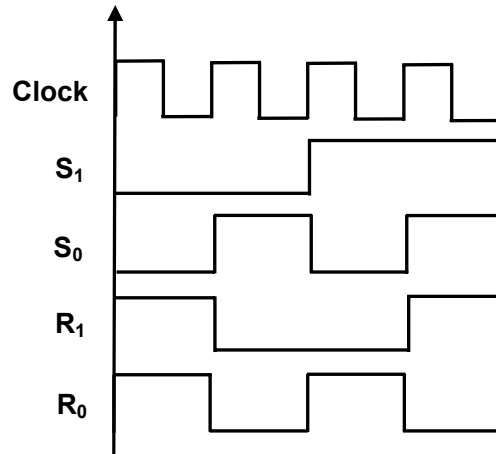


Figure 6.17: Clock and control signals for SDR driver-receiver cores.

For the DDR mechanism discussed in the previous subsections, the clock signals can be directly used to generate S_0 and R_0 , and a toggle latch is used for generating S_1 and R_1 . However, as seen in Figure 6.17, with the SDR approach, a toggle latch will be needed for generating of S_0 and R_0 , and a 2-stage toggle latch can be used for generating S_1 and R_1 . The circuit diagram for the generalized 2-stage toggle latch is shown in Figure 6.18. The first stages for the actual toggle latch circuits corresponding to two clocks are slightly different from each other because of their triggering mechanisms. The timing

diagram for SDR driver-receiver cores is shown in Figure 6.19 corresponding to a random phase difference between CLK_S and CLK_R.

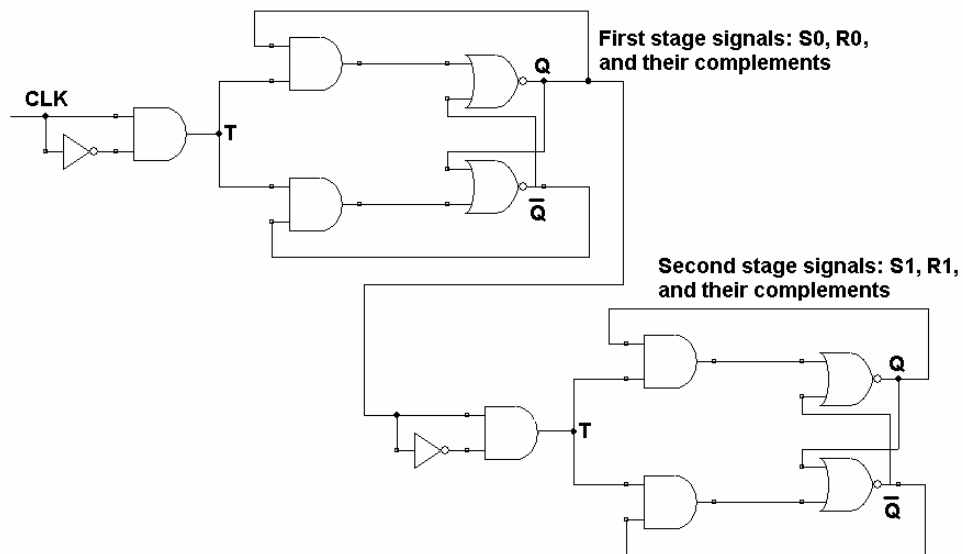


Figure 6.18: A generalized 2-stage toggle latch for generating control signals for SDR driver-receiver cores in GALS scenario.

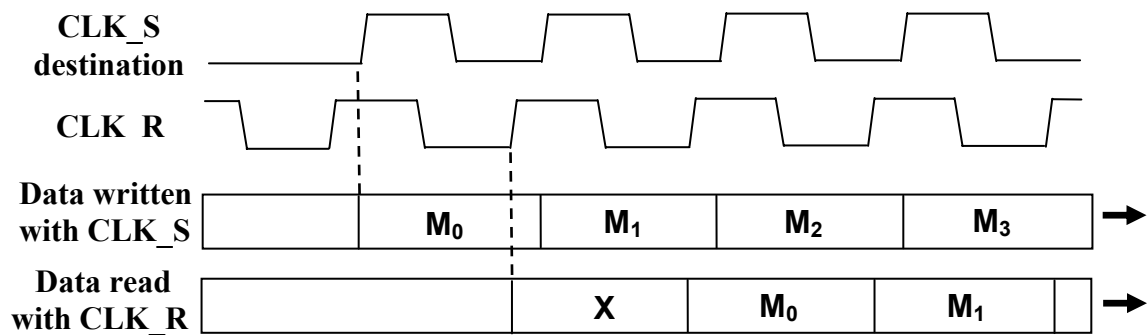


Figure 6.19: Waveforms for retimer circuit for SDR driver-receiver cores.

Thus, it is shown in this section that wave-pipelining is an effective technique for SoC global interconnects, which use GALS interfaces between various cores. The retimer circuit discussed in this section facilitates latching of the data in a clock domain different from the source domain, by a locally generated clock, which can be skewed to any extent w.r.t. the source clock.

6.6 Summary

In this chapter, wave-pipelining is compared to latch insertion under a constant throughput constraint and is shown to give significantly better results for all of latency, power, and silicon area. A brief overview of the timing circuits proposed by some researchers to synchronize the data on wave-pipelined interconnect circuits is presented in this chapter, along with a discussion of the advantages and shortcomings of these circuits. A new receiver for wave-pipelined interconnects is constructed by modifying the existing circuits to overcome their shortcomings.

To aid the communication between multiple clock domains in an SoC using the GALS interface, a new receiver circuit is proposed, which can synchronize the data to a locally generated receiver clock that has a completely random phase w.r.t. the source clock. The performance and robustness of this receiver circuit is verified through HSPICE simulations. Wave-pipelining is shown to be a very effective technique for high-performance global interconnects in fully synchronous or globally asynchronous locally synchronous VLSI systems.

CHAPTER 7

IMPACT OF TECHNOLOGY SCALING ON WAVE-PIPELINING

7.1 Introduction

The circuit- and system-level analysis of the wave-pipelined interconnects is performed in the previous chapters, and a 180 nm technology node is used for HSPICE simulations and validation. However, to understand the usefulness of wave-pipelining for future processor generations, it is important to study the impact of technology scaling on wave-pipelining. The analytical throughput model derived in Chapter 2 is general and can be applied to the interconnects of any technology generation, therefore, this chapter uses the predictive capability of the analytical throughput model to study the future of wave-pipelining.

Based on the ITRS roadmap [2], eight different technology generations between 180 nm and 18 nm are considered in this chapter, and the performance of wave-pipelined interconnect circuits in these technology generations is predicted using the performance models derived in Chapter 2. The important technology parameters for these technology nodes are listed in Table 7.1.

Table 7.1: Technology parameters for different technology generations [2].

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Physical gate length (nm)	140	90	53	32	22	16	11	9
Supply Voltage (V)	1.8	1.2	0.9	0.8	0.7	0.6	0.5	0.5
Threshold Voltage (V)	0.42	0.30	0.26	0.26	0.22	0.21	0.17	0.17

First, the evaluation of the transistor and interconnect parameters for future technology generations is discussed. The analytical throughput model in (2.16), which is used as the primary tool in this chapter to project the performance of future interconnects, is then validated for a 45 nm node with HSPICE simulations using the transistor models in [76]. This is followed by a detailed analysis of the impact of technology scaling on performance of wave-pipelined global interconnects. A 32 nm node is chosen as a representative example of the future technology generation, and its complete performance, power, and area analysis is performed. Finally, the impact of various material constants and technology-dependent parameters on the performance of future interconnect circuits is discussed.

7.2 Evaluation of transistor parameters

7.2.1 Evaluation of transistor resistance

The first order approximation for the transistor resistance, R_t , is given in [6] as

$$R_t \approx \frac{1}{\mu C_{ox} \left(\frac{W}{L} \right) (V_{dd} - |V_t|)} \quad (7.1)$$

The term C_{ox} is given by

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}, \quad (7.2)$$

where ϵ_{ox} is the permittivity of the oxide and t_{ox} is its thickness. Using (7.2), (7.1) can be further expanded as

$$R_t \approx \frac{1}{\mu \frac{\epsilon_{ox}}{t_{ox}} \left(\frac{W}{L} \right) (V_{dd} - |V_t|)}. \quad (7.3)$$

For constant field scaling, assuming ϵ_{ox} remains constant, the scaling of W and L cancel each other, and that of t_{ox} and the voltages cancel each other. Therefore, it can be seen from (7.3) that R_t depends only on carrier mobility μ . ITRS has provided a mobility enhancement factor for every technology generation. Applying this factor to R_t for the 180 nm node, whose evaluation is shown in Appendix, R_t for every future technology generation is calculated. Table 7.2 shows the values of R_t and other transistor parameters corresponding to a transistor scaling factor of 56, which is used in earlier chapters. The base value of 180 ohm for the 180 nm technology is highlighted in Table 7.2.

Table 7.2: Transistor resistance and other parameters for a transistor scaling factor of 56, for different technology generations.

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
NMOS width (nm)	7840	5040	2968	1792	1232	896	616	504
PMOS width (nm)	19600	12600	7420	4480	3080	2240	1540	1260
Mobility enhancement factor	1	1	1	1	1.3	2	2	2
R_t (ohm)	180	180	180	180	138	90	90	90

It is seen from Table 7.2 that R_t decreases with technology scaling because of mobility enhancement. However, it is important to note that though an enhancement in mobility is projected for the future technology generations, the manufacturing solutions to achieve it are not completely known [2]. If this enhancement in mobility is not achieved, R_t could remain almost unchanged despite technology scaling.

7.2.2 Evaluation of transistor capacitance

The primary contribution to the transistor capacitance, C_t , comes from the gate capacitance [47], therefore, C_t can be approximated as

$$C_t \approx \frac{\epsilon_{ox}}{t_{ox}}(W_p + W_n)L, \quad (7.4)$$

where W_p and W_n denote the widths of a PMOS and NMOS, respectively. For constant field scaling, assuming ϵ_{ox} remains constant, the scaling of t_{ox} and the widths cancel each other, and the scaling of C_t is then proportional to the channel length scaling. Using the base value for C_t for a 180 nm transistor with a scaling factor of 56, the values of C_t are presented in Table 7.3 for different technology generations. The base value of 130 fF for the 180 nm technology node is highlighted in Table 7.3.

Table 7.3: Transistor capacitance for a transistor scaling factor of 56, for different technology generations.

Technology Generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
C_t (fF)	130	94	65	47	32	23	16	13

7.3 Evaluation of interconnect parameters

The evaluation of interconnect resistance, R , capacitance, C , and inductance, L , is discussed in this section. To determine these parameters, it is important to know the interconnect dimensions for future technology generations. Because the focus of this research is on the global interconnects, the predictions are made for global interconnect parameters.

Intel has released the process data for technology nodes up to 65 nm [77], which contain a detailed information about the interconnect dimensions. Therefore, Intel interconnects dimensions are used up to the 65 nm node, and these data points are extrapolated to predict the global interconnect dimensions for future technology generations. It is observed in [29], [77]-[79] that despite technology scaling, the global interconnect dimensions have been fairly constant. Therefore, it is assumed that the generations beyond 65 nm will have a global interconnect pitch of almost 1000 nm on the topmost layer, which is also consistent with the ITRS projections [2]. The values of global wire pitch are shown in Figure 7.1 to show the trend in the Intel technology and its extrapolation for future generations.

The values for the interconnect height-to-width aspect ratio (AR wire) and the thickness-to-width ratio (AR via) used in this chapter are based on the ITRS roadmap [2], which are also consistent with Intel's released process data for up to the 65 nm node. The dimensions for a global interconnect on the topmost metal layer are presented in Table 7.4, which are used to calculate the interconnect parameters in the following sections.

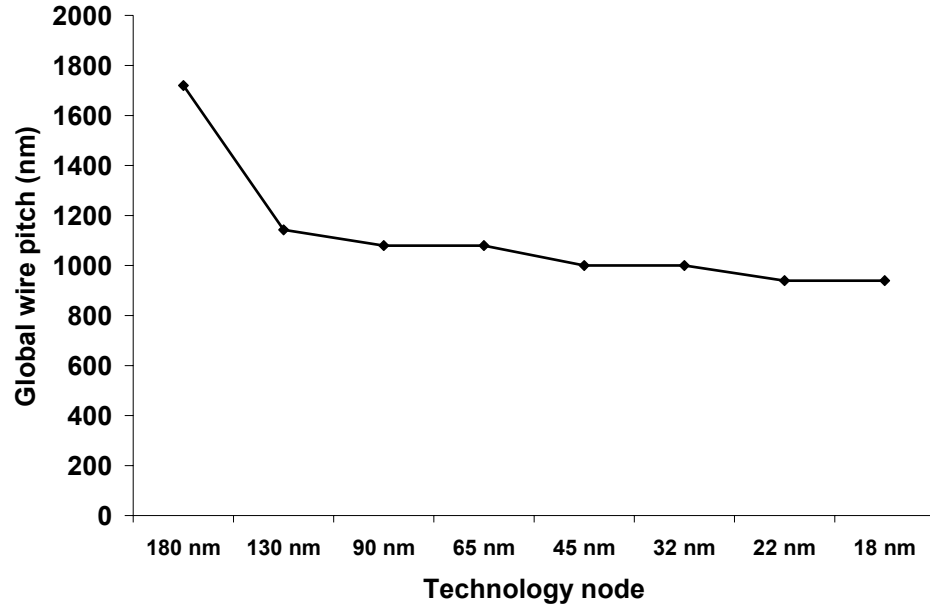


Figure 7.1: Values of global wire pitch for present technology generations and their extrapolation for future technology generations.

Table 7.4: Global interconnect dimensions for different technology generations.

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Global pitch (nm)	1720	1143	1080	1080	1000	1000	940	940
AR wire	2	2.1	1.8	1.8	1.8	1.8	1.8	1.8
AR via	1.8	1.8	1.6	1.6	1.6	1.6	1.6	1.6
w (nm)	860	570	540	540	500	500	470	470
s (nm)	860	573	540	540	500	500	470	470
h (nm)	1720	1200	972	975	900	900	846	846
t (nm)	1548	1026	864	864	800	800	752	752

7.3.1 Evaluation of interconnect resistance

The interconnect resistance, R , can be expressed as

$$R = \frac{\rho l}{wh} \quad (7.5)$$

Because of severe scattering, ITRS has projected the resistivity (ρ) to be 3.6 $\mu\text{ohm-cm}$ for the local interconnects of the 18 nm generation. However, because of larger interconnect

dimensions, ρ for the global interconnects of future technology generations is expected to be equal to its value of 2.2 $\mu\text{ohm-cm}$ for the 180 nm generation [2]. Therefore, using ρ of 2.2 $\mu\text{ohm-cm}$ and the interconnect dimensions in Table 7.4, the values of R for a 1 cm interconnect are shown in Table 7.5 for different technology generations between 180 nm and 18 nm.

Table 7.5: Interconnect resistance per unit cm for different technology generations.

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Resistance of a 1 cm long interconnect (ohm)	148.73	321.64	419.14	417.85	488.89	488.89	553.41	553.41

7.3.2 Evaluation of interconnect capacitance

Using the values for the relative permittivity of the dielectric (ϵ_r) from [2] and the interconnect dimensions in Table 7.4, the interconnect capacitance, C , is calculated using RAPHAEL. A 5-interconnect, 2- ground plane system is considered in which the two interconnects at far ends are ground lines. The values of C are shown in Table 7.6 for a 1 cm long interconnect. It is seen that C reduces with technology scaling, and this reduction primarily comes from the reduction of ϵ_r .

Table 7.6: Interconnect capacitance per unit cm for different technology generations.

Technology Generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Relative permittivity of dielectric (ϵ_r)	3.5	3.3	3.1	2.7	2.3	2	1.9	1.8
Capacitance of a 1 cm long interconnect (pF)	2.256	2.191	1.927	1.678	1.429	1.243	1.181	1.119

7.3.3 Evaluation of interconnect inductance

The inductance L of a 1 cm long global interconnect is calculated using RAPHAEL for a 5-interconnect system, in which two interconnects at far ends act as the ground lines and provide the necessary return paths. The values of the self inductance and the coupling factors between neighbors are shown in Table 7.7 for the interconnect dimensions shown in Table 7.4. The inductance coupling factor is defined as

$$\text{Inductance coupling factor} = \frac{\text{Mutual inductance } (L_m)}{\text{Self inductance } (L_s)}. \quad (7.6)$$

It is seen in Table 7.7 that technology scaling does not significantly affect the self and mutual inductance of global interconnects.

Table 7.7: Interconnect inductance per unit cm and coupling factors for different technology generations.

Technology Generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Inductance of a 1 cm long interconnect (nH)	3.215	3.053	3.295	3.295	3.289	3.289	3.289	3.289
Coupling factor between near neighbors	0.5233	0.5155	0.5102	0.5102	0.5096	0.5096	0.5096	0.5096
Coupling factor between far neighbors	0.3010	0.2916	0.2891	0.2891	0.2886	0.2886	0.2886	0.2886

7.4 Comparison of throughput using analytical expression and HSPICE for a 45 nm node

The comparison of throughput using the analytical model in (2.16) and HSPICE simulations is presented in Section 2.3.2 for a 180 nm node, and this section tries to analyze the applicability and accuracy of the analytical throughput model for future

technology generations. The Berkeley Predictive Transistor Model (BPTM) provides HSPICE transistor parameters for some future technology generations [76]. The smallest technology generation for which the parameters are available is 45 nm, therefore, this 45 nm node is selected to validate the analytical throughput model with HSPICE simulations.

A 1 cm long 45 nm wave-pipelined global interconnect is used for this analysis. The physical parameters used for HSPICE simulations are shown in Table 7.8. Based on the transistor parameters in Table 7.8, R_t of 522 ohm and C_t of 32.5 fF are used to evaluate the throughput of the wave-pipelined interconnect using (2.16).

Table 7.8: Physical parameters used for 45 nm HSPICE simulations.

Technology size	45 nm
Interconnect parameters (cross sectional dimensions = 500 nm x 900 nm)	
Resistance	489 ohm/cm
Capacitance	1.429 pF/cm
Inductance	3.289 nH/cm
Repeater driver parameters (level-54 models)	
pMOS	Length = 0.045 μm and Width = 2.52 μm
nMOS	Length = 0.045 μm and Width = 6.30 μm

Figure 7.2 shows the variation of throughput with the number of repeaters per cm using the analytical model in (2.16) and HSPICE simulations. It is seen in Figure 7.2 that the values of throughput calculated using the analytical expression fairly match with those from HSPICE simulations. The comparison of Figure 2.7, which shows the throughput for 180 nm interconnects, and Figure 7.2 shows that the trend of variation of the throughput with the repeater density remains unchanged despite technology scaling. Therefore, the analytical throughput model can be effectively used to correctly predict the throughput of the wave-pipelined interconnects of future technology generations.

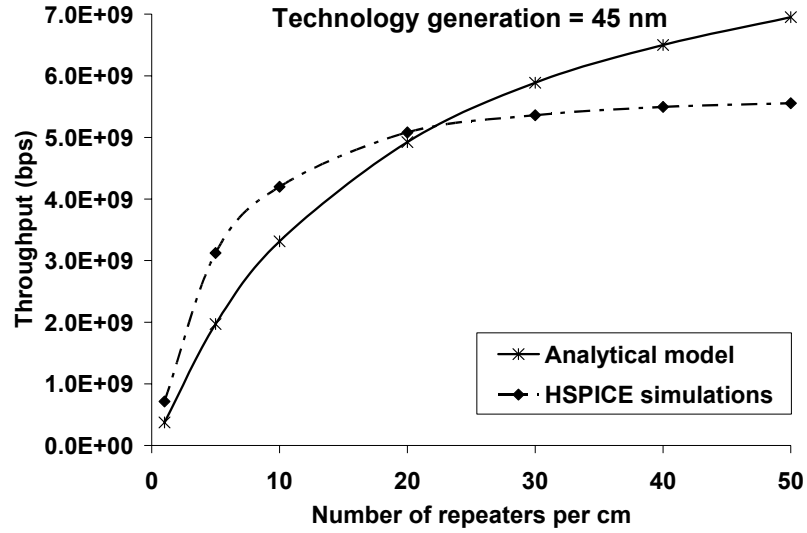


Figure 7.2: Comparison of throughput using analytical model and HSPICE simulations for a 1 cm long 45 nm interconnect.

Based on HSPICE simulations for the 45 nm node, the values of latency, the corresponding reciprocal latency bit rate, and the throughput bit rate are shown in Table 7.9. The results in Table 7.9 show that wave-pipelining continues to give a significantly higher throughput (larger by an order of magnitude) compared to the reciprocal of latency for future technology generations, which underlines the importance of wave-pipelining for the performance enhancement of future global interconnects.

Table 7.9: Comparison of bit rates using reciprocal latency and throughput for 45 nm global interconnects.

Number of repeaters per cm	50% Latency (ns)	Bit rate (Gbps)	
		Reciprocal latency	Throughput
1	0.62	1.603	0.714
5	0.79	1.266	3.125
10	1.07	0.935	4.200
20	1.63	0.614	5.080
30	2.21	0.453	5.360
40	2.77	0.362	5.495
50	3.32	0.301	5.556

7.5 Impact of technology scaling on communication throughput

7.5.1 ITRS projections for on-chip local clock frequency

The values of the on-chip local clock frequency projected by ITRS for upcoming technology generations [2] are shown in Table 7.10. Because the interconnects are the primary performance bottlenecks, these values, which are a measure of performance, must be achieved on the interconnects. Even though the focus is on global interconnects, a current trend toward the globally asynchronous locally synchronous (GALS) systems suggests that these interconnects could be pipelined and run at the on-chip local clock frequency. Therefore, a performance analysis of the wave-pipelined interconnects is performed in this section for the technology generations up to 18 nm to see if future interconnects achieve a throughput that is comparable to the on-chip local clock frequency.

Table 7.10: On-chip local clock frequency for different technology generations [2].

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
On-chip local clock frequency (GHz)	1.20	1.68	4.17	9.29	15.08	22.98	39.68	53.21

The timing and synchronization analysis of wave-pipelined interconnect circuits in Chapter 6 shows that if the data on wave-pipelined interconnects are latched at the output by sending the clock along with data, a 53 GHz clock frequency for the 18 nm node could translate into a 106 Gbps throughput on the clock line. However, it should be noted that a 53 GHz clock frequency is projected by ITRS to process or transfer *data* at the rate of 53 Gbps. If the double data rate (DDR) approach is used, a 53 Gbps data transfer rate is obtained by sending a 26.5 GHz (or 53 Gbps) clock on the interconnect.

At the receiver cores, the data can be synchronized to a 53 GHz clock if needed. Because the focus of this analysis is on *data* interconnects, the target throughput is assumed to be equal to the clock frequency for simplicity.

7.5.2 Impact of technology scaling on throughput for two different design scenarios

To study the impact of technology scaling on interconnect throughput, two different design scenarios, which use different supply voltages, are considered. Figure 7.3 shows the variation of throughput with repeater density for different technology generations for two supply voltages, the typical supply voltage used for that technology node and a supply voltage equal to twice the threshold voltage, as suggested in Chapter 5 for the low-power design. The optimal repeater size h_{opt} , which is calculated using (2.41), is used to design the repeater circuits for every technology generation. The clock frequency for every technology generation is also shown in Figure 7.3 by dotted lines to obtain the values of the repeater density needed to meet the projected throughput. It is seen in Figure 7.3 that the projected values of throughput are easily met on the 1 cm long interconnect even when the scaled supply voltage of $2|V_t|$ is used. Therefore, VSRI is a viable technique for present as well as future technology generations.

As a side note, a keen observation of Figure 7.3 shows that by operating with the typical supply voltages and optimal-sized repeaters, the maximum values of throughput obtained are almost *twice* the clock frequency. Therefore, the clock lines can also be successfully routed using the same dimensions as data lines. For instance, Figure 7.3 shows that for the 18 nm technology node, if the 53 GHz clock is sent along with data, the required 106 Gbps clock throughput can be easily achieved on the wave-pipelined

global interconnect. However, because the focus is on data interconnects, the target throughput for the 18 nm node is assumed to be 53 Gbps.

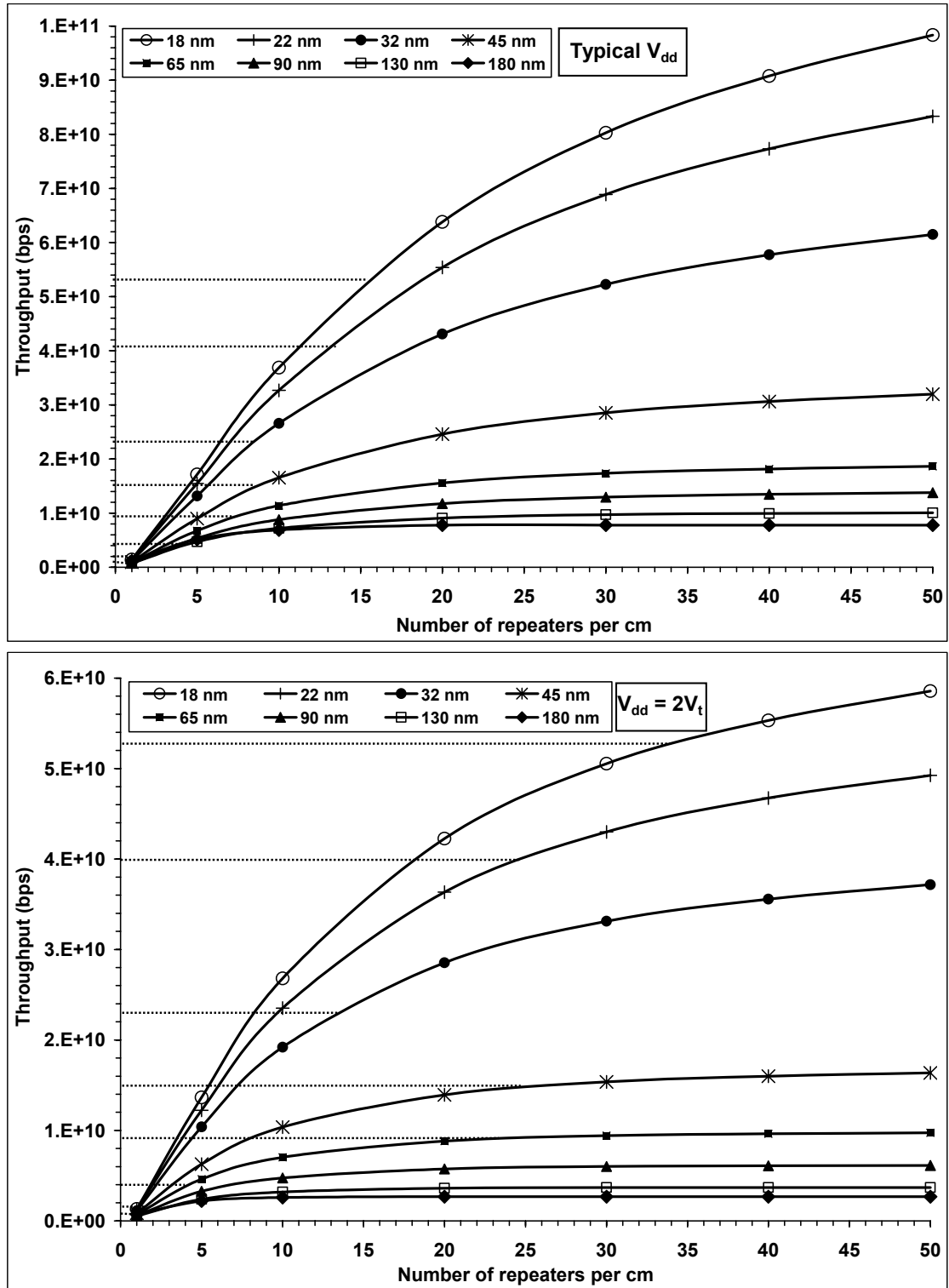


Figure 7.3: Throughput using optimal-sized repeaters for different technology generations.

7.5.3 Future of latency-centric repeater insertion

To analyze the impact of technology scaling on latency-centric repeater insertion, a 1 cm long global interconnect, which operates with the typical supply voltage and uses the optimal number and size of repeaters in [6], is considered. The values for the optimal number and size of repeaters and the corresponding values of the optimal latency are shown in Table 7.11. Figure 7.4 compares the optimal reciprocal latency to the target clock frequency projected by ITRS. It is seen in Figure 7.4 that the values of throughput by latency-centric repeater insertion are up to an *order of magnitude smaller* than those projected by ITRS for future technology generations.

Table 7.11: Optimal values of parameters for latency-centric repeater insertion on a 1 cm long interconnect.

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
Optimal number of repeaters (n_{opt}) [6]	2	4	5	6	8	11	13	14
Optimal size of repeaters (h_{opt}) [6]	256	202	200	219	197	176	207	222
Optimal latency (ns)	0.673	0.829	0.739	0.585	0.426	0.271	0.218	0.192

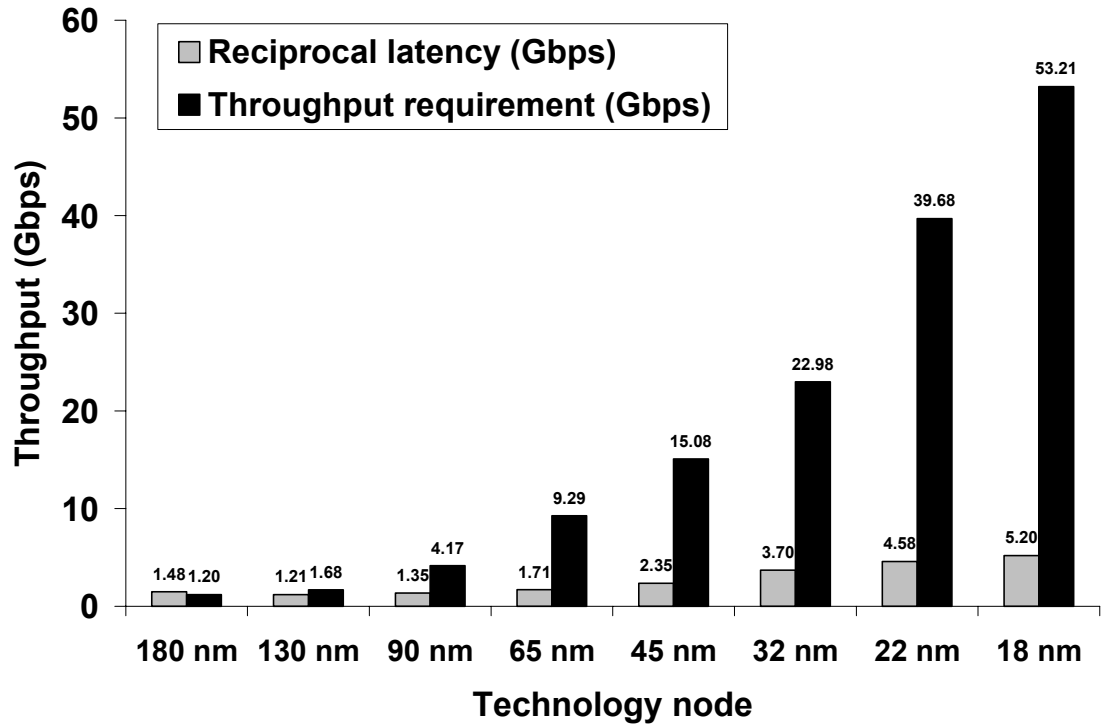


Figure 7.4: Comparison of reciprocal latency and throughput requirement for various technology generations.

7.5.4 Impact of technology scaling on power and performance of global interconnects

It is clear from Figure 7.4 that latency-centric repeater insertion falls significantly short of the ITRS-projected throughput for the future technology nodes. Wave-pipelining, on the other hand, is shown to meet these projected throughput requirements in Figure 7.3. To study the performance and power trends in wave-pipelined global interconnects, a detailed analysis of 1 cm long global interconnect is performed for all technology nodes up to 18 nm. It is assumed that the wave-pipelined interconnect operates with the typical supply voltage and uses optimal-sized repeaters. The power and latency of the 1 cm long interconnect, required to meet the projected values of throughput, are shown in Table

7.12. The switching power (dynamic + short-circuit) is calculated assuming an activity factor of 0.1 [28], [34].

Table 7.12 shows that multiple clock cycles are needed for data transmission in future technology generations, and single cycle operation may not be feasible on future global interconnects. It should also be noted from Table 7.12 that the switching power is projected to increase by a factor of 2 from 180 nm to 18 nm, whereas the leakage power is projected to increase by a factor of 500. The leakage power, which is 0.2% of the total power for the 180 nm node, is 34% of the total power for the 18 nm node. Therefore, leakage power is a significant source of power dissipation for future technology generations, and the techniques to reduce leakage power need to be employed in future systems.

Table 7.12: Power and performance of a 1 cm long global interconnect to achieve the projected throughput for various technology generations.

Technology generation	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	18 nm
On-chip local clock frequency (GHz)	1.20	1.68	4.17	9.29	15.08	22.98	39.68	53.21
Optimal repeater size (h_{opt})	256	202	200	219	197	176	207	222
Number of repeaters per cm	2	3	5	7	8	8	14	16
Latency in clock cycles	0.81	1.43	3.10	5.54	6.46	6.29	8.71	10.30
Switching power (mW)	0.947	0.538	0.753	1.443	1.355	1.108	1.472	1.938
Subthreshold + gate leakage power (mW)	0.002	0.008	0.208	0.339	0.348	0.533	0.983	1.002

7.5.5 Impact of technology scaling on supply voltage

It would be interesting to compare the projected supply voltages for various technology generations to the optimal supply voltage of $2|V_t|$, which is suggested in this research for low-power applications. The typical supply voltages and threshold voltages for various technology generations are listed earlier in Table 7.1, and Figure 7.5 shows a comparison between the *ratios* of the supply voltage to the threshold voltage suggested by ITRS [2] and that proposed in (5.10), for various technology generations. It is seen in Figure 7.5 that this ratio decreases with technology scaling, which indicates that the ITRS roadmap scales down the supply voltage more aggressively than the threshold voltage. However, even for an 18 nm technology node, ITRS projects this ratio to be more than two, which is the optimal ratio suggested in this work. Thus, this research suggests more aggressive supply voltage scaling than ITRS for high-performance, low-power interconnect circuits.

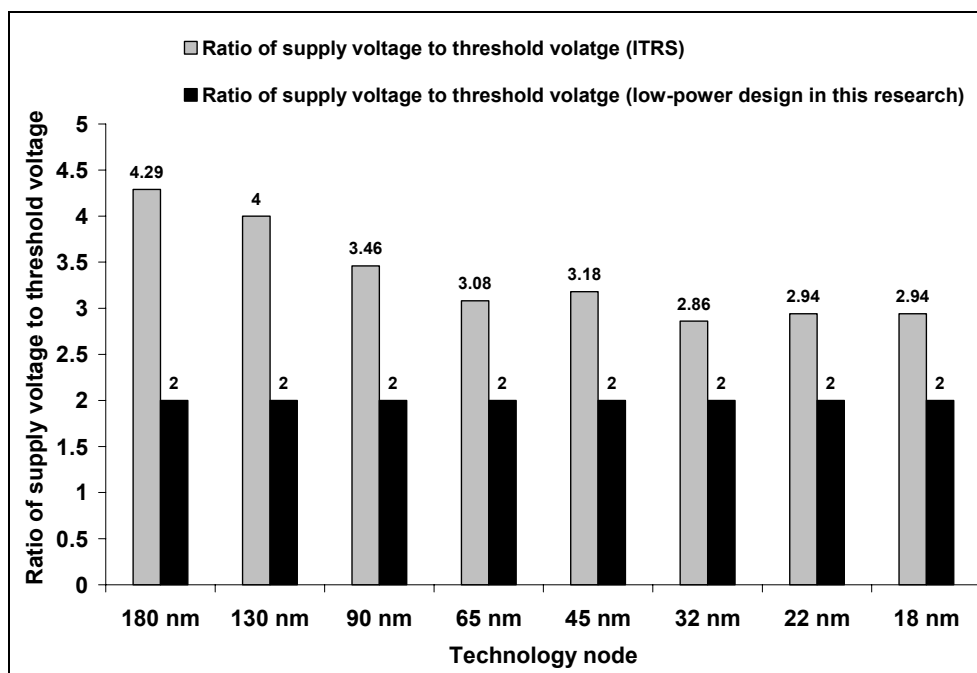


Figure 7.5: Comparison of ratios of supply voltage to threshold voltage by ITRS [2] and low-power design in this research, for different technology generations.

7.6 Performance, power, and area analysis of 32 nm node

The year of production for the 32 nm technology node is projected to be 2013 [2]. The process parameters for this node are not expected to be released in the near future. However, the manufacturing solutions to achieve the projected values for some of the parameters such as the mobility enhancement ratio, effective dielectric constant, effective resistivity, etc. are known for this technology node. Therefore, the 32 nm node is chosen to perform a holistic analysis and estimate the performance, power, and area for the future interconnect circuits. Based on the ITRS roadmap [2], the projected dimensions and other parameters for an interconnect architecture that has 12 metal levels are shown in Table 7.13.

Table 7.13: A 32 nm interconnect system with 12 metal levels.

Metal level	Pitch (nm)	AR wire	AR via	ρ ($\mu\text{ohm-cm}$)	ϵ_r
1	76	1.9	1.7	3.04	2
2	76	1.9	1.7	3.04	2
3	84	1.9	1.7	3.04	2
4	86	1.9	1.7	3.04	2
5	96	1.9	1.7	3.04	2
6	120	1.9	1.7	3.04	2
7	180	2.0	1.8	2.8	2
8	260	2.0	1.8	2.8	2
9	380	2.2	2.0	2.2	2
10	520	2.2	2.0	2.2	2
11	700	2.2	2.0	2.2	2
12	1000	2.2	2.0	2.2	2

7.6.1 Design optimization for 32 nm node

Four different design choices that are considered for the design of a 1 cm long wave-pipelined global interconnect on metal-12 are listed below.

A. Typical V_{dd} and $h = 56$

B. $V_{dd} = 2|V_t|$ and $h = 56$

C. Typical V_{dd} and $h = h_{opt}$

D. $V_{dd} = 2|V_t|$ and $h = h_{opt}$

The first two designs use a repeater scaling factor of 56 (which is used in the previous chapters in this thesis), to analyze the impact of suboptimal repeater sizes on performance, power, and area. The 1 cm long interconnect is designed to achieve a throughput of 22.98 Gbps, which is equal to the on-chip local clock frequency for this technology node. The design parameters for these four design choices are shown in Table 7.14.

Table 7.14: Design parameters for a 1 cm long 32 nm global interconnect routed on metal-12.

Design	A	B	C	D
Supply voltage (V)	0.6	0.42	0.6	0.42
Repeater size	56	56	176	240
Number of repeaters	14	32	8	14
Throughput (bps)	2.298E+10			
R (ohm)	488.89	488.89	488.89	488.89
C (F)	1.243E-12	1.243E-12	1.243E-12	1.243E-12
R_t (ohm)	90	167.14	28.62	38.98
C_t (F)	2.311E-14	2.311E-14	7.272E-14	9.910E-14

The power and area for all the design choices are shown in Table 7.15. The wire area and silicon area are fairly simple to calculate because all the necessary transistor and interconnect dimensions are known. The total power shown in Table 7.15 includes all of the dynamic, short-circuit, and leakage power. The worst-case dynamic power is calculated using (4.1) and the short-circuit power is calculated using its model in [12].

The value of the static power per micron transistor width in the ITRS roadmap [2] is used to calculate the leakage power for repeater drivers. This value includes both the subthreshold leakage and the gate leakage. Based on the values of power and area, the values of the TPBE, TPEA, and TPA metrics, which are discussed in Chapter 5, are calculated. The optimal values of these metrics are highlighted in Table 7.15.

Table 7.15: Performance, power, and area for a 1 cm long 32 nm global interconnect.

Design	A	B	C	D
Throughput (bps)	2.298E+10			
Wire area (cm ²)	1E-04	1E-04	1E-04	1E-04
Silicon area (cm ²)	0.983E-07	2.248E-07	1.766E-07	4.215E-07
Total bit energy (pJ)	0.348	0.249	0.448	0.372
Total power (mW)	7.999	5.728	10.300	8.539
TPBE (bps/pJ)	6.602E+10	9.219E+10	5.129E+10	6.185E+10
TPEA (bps/pJ.cm ²)	6.595E+14	9.199E+14	5.120E+14	6.159E+14
TPA (bps/cm ²)	2.296E+14	2.293E+14	2.294E+14	2.288E+14

It is seen in Table 7.15 that Design B, which uses a scaled supply voltage and $h = 56$, results in minimal bit energy, which maximizes both TPBE and TPEA for the given throughput. This fact underlines the importance of voltage scaling for the future technology generations. Design A, which uses the typical supply voltage and $h = 56$, results in the maximum TPA among the four designs considered in Table 7.15. It is also seen from the results in Table 7.15 that *though optimal repeater sizing maximizes the throughput, it is not necessarily area- and power-efficient*. As seen in Table 7.15, using a slightly larger number of suboptimal-sized repeaters can achieve the same throughput at the expense of less silicon area and power.

7.6.2 Impact of via area on design optimization

The insertion of repeaters increases the number of vias in a system architecture [32]. The area occupied by the vias reduces the wiring efficiency of a metal layer. Optimal wire sizing to reduce the impact of the number of repeaters on via blockage is discussed earlier in Section 5.7. The impact of via area on the optimal TPEA and TPA designs for the 32 nm node is analyzed in this subsection. Using the basic via model in [32], the via area for all the design choices in Table 7.14 is shown in Table 7.16. The total area is now given by the sum of the silicon area, wire area, and via area, which is used to recalculate the area-dependent metrics, TPEA and TPA. The optimal values of TPEA and TPA are highlighted in Table 7.16.

Table 7.16: Impact of via area on design metrics.

Design	A	B	C	D
Via area (cm ²)	5.752E-07	13.152E-06	3.287E-07	5.752E-07
TPEA (bps/pJ.cm ²)	6.557E+12	9.080E+12	5.103E+12	6.124E+12
TPA (bps/cm ²)	2.283E+14	2.263E+14	2.286E+14	2.275E+14

A comparison between Table 7.15 and Table 7.16 shows that the optimal TPEA design point is unchanged by the inclusion of the via area to the metric, but the optimal TPA design point changes from Design A to Design C. It is seen from this particular example that the use of optimal-sized repeaters, which reduces the required number of repeaters thereby proportionally reducing the via area, could be a better design option than suboptimal-sized repeaters for certain area-centric applications. This example also shows that the sensitivity of a particular design method to the different underlying parameters must be clearly understood before it is optimized.

7.6.3 Power breakdown for a 32 nm global interconnect circuit

The breakdown of the total power into its different components is shown in Figure 7.6 for Design A, which uses a typical supply voltage of 0.6 V and $h = 56$. The pie chart shown in Figure 7.6 assumes an activity factor of 0.1 [28], [53]. It is seen in Figure 7.6 that almost 85% of the total power is the dynamic power, and the remaining portion is the leakage power. Compared to the earlier technology generations, the leakage power for a 32 nm node is a larger fraction of the total power [12], which is a result of a smaller threshold voltage, a larger subthreshold leakage current, and a smaller gate oxide thickness [2]. This fact is also supported by the results in Table 7.12. The short-circuit power for this technology node is negligible because the time for which both the transistors remain simultaneously ON is extremely small as a result of a smaller supply voltage and a smaller supply-to-threshold ratio.

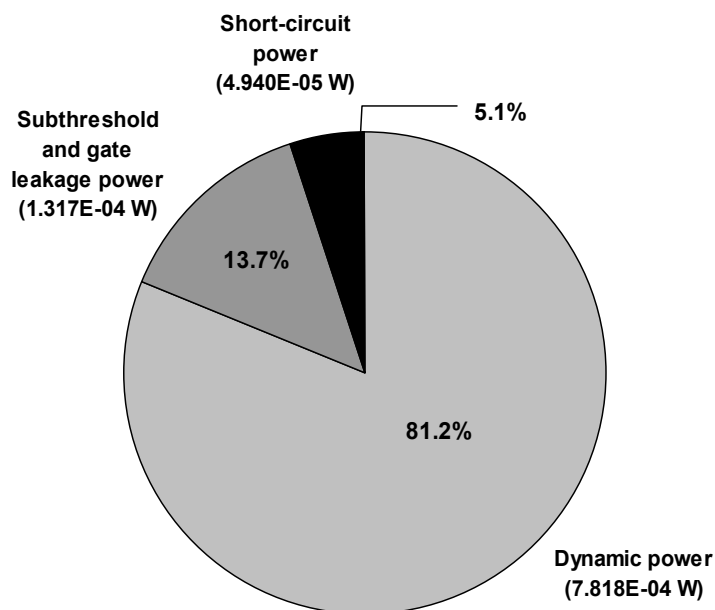


Figure 7.6: Power breakdown for a 32 nm global interconnect circuit.

7.6.4 Multilayer interconnect throughput for 32 nm node

To assess the performance of different metal layers in an interconnect architecture, the communication throughput for different metal layers is calculated using the interconnect dimensions shown in Table 7.13. This analysis is performed for two different interconnect lengths, 1 mm and 1 cm. Figure 7.7 shows the throughput of a 1 cm long interconnect for all 12 metal layers of the 32 nm node, as the number of repeaters per unit cm varies from 1 to 50, whereas Figure 7.8 shows the throughput of a 1 mm long interconnect as the number of repeaters per unit mm varies from 1 to 5. In Figure 7.7 and Figure 7.8, the reference lines corresponding to the on-chip local clock frequency of 23 GHz are shown to find out which metal layers meet this projection.

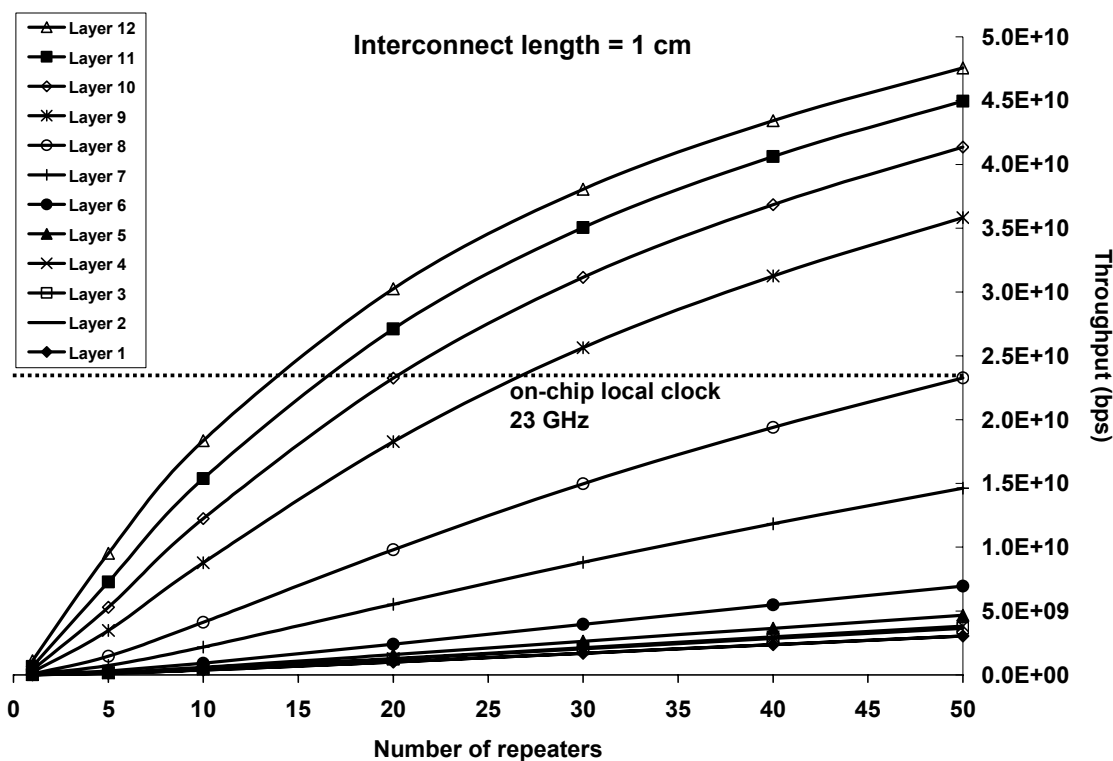


Figure 7.7: Throughput of a 1 cm interconnect in 32 nm technology generation.

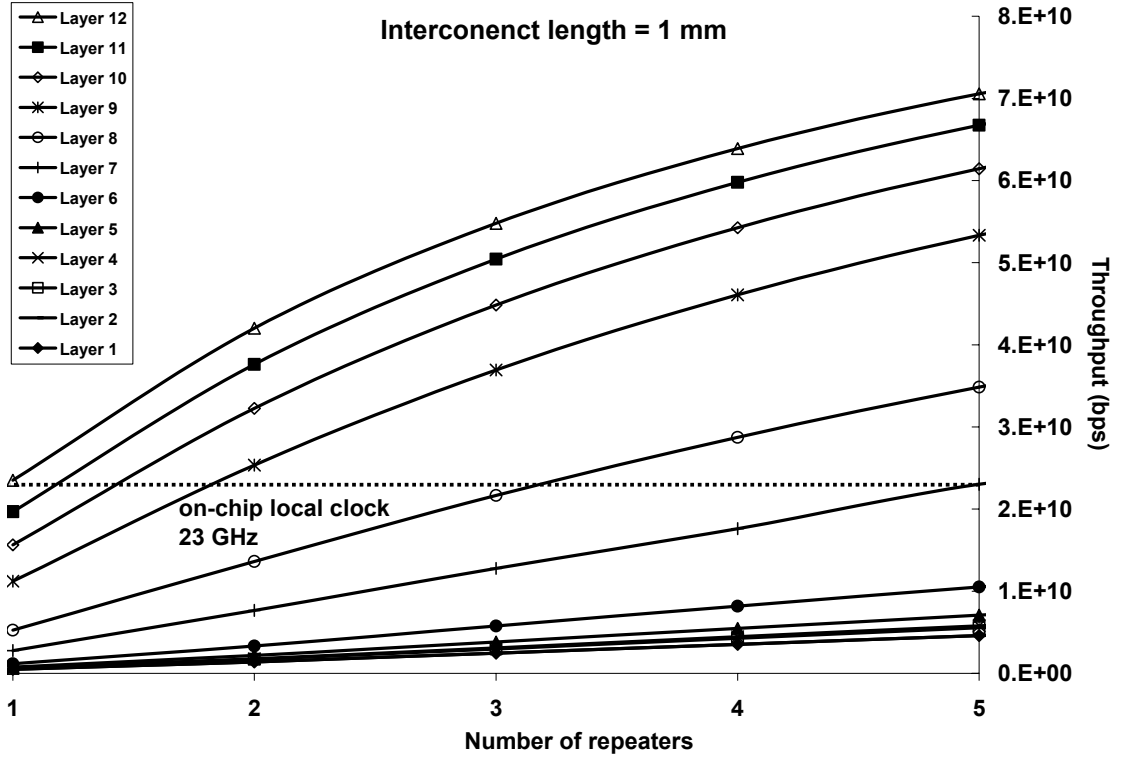


Figure 7.8: Throughput of a 1 mm interconnect in 32 nm technology generation.

The repeater density of 50 repeaters per cm in Figure 7.7 translates into 5 repeaters per mm for Figure 7.8. However, the throughput achieved on a 1 mm long interconnect is significantly higher than that achieved on a 1 cm long interconnect for this same repeater density. As discussed earlier in Section 2.4.1, for the same repeater density, the first segment of a smaller interconnect needs to undergo a smaller voltage swing, which results in a higher throughput.

It is seen in Figure 7.7 that a 23 Gbps throughput is achieved on a 1 cm long interconnect on metal-8 and above, whereas it is achieved on metal-7 and above on a 1 mm long interconnect. It is seen that as the interconnect length decreases, the required throughput is also achievable on the lower metal levels, which would typically route these smaller interconnects. This fact can be extrapolated to conclude that the required on-chip local clock frequency can be met on most interconnects of the 32 nm node.

7.7 Impact of material alternatives on wave-pipelining

It is seen in the earlier sections that the performance predicted by the ITRS roadmap [2] for future technology generations can be achieved if the values for the underlying parameters are reached. However, the manufacturing solutions for obtaining low dielectric permittivity, high mobility for carriers, etc., which have a significant impact on the throughput, are not yet known [2] for some of the future technology nodes. This section analyzes the impact of such parameters on the throughput performance. The difference between the projected performance and the actual performance if the projected values of the underlying parameters are not achieved is discussed in this section using a 32 nm node.

7.7.1 High resistivity resulting from scattering

As a result of an increased scattering for the interconnects with smaller dimensions, the resistivity, ρ , of the metal is significantly high for the lower (local) metal levels, as seen in Table 7.13. As a result, the interconnects on lower levels have a very high resistance. For the 32 nm interconnect system shown in Table 7.13, a 1 mm interconnect on metal-6 that has a 120 nm pitch and ρ of 3.04 $\mu\text{ohm-cm}$ is analyzed in this subsection. The resistance and capacitance parameters are shown in Table 7.17.

Table 7.17: Values of resistance and capacitance for a 1 mm long 32 nm metal-6 interconnect.

R	C	R_t	C_t
4444.44 ohm	127 ff	90 ohm	23.11 ff

It is seen in Figure 7.8 that if this interconnect is wave-pipelined by inserting 5 repeaters per mm, a throughput of 10.5 Gbps can be obtained. To achieve a 23 Gbps throughput (corresponding to the on-chip local clock frequency of 23 GHz), 12 repeaters need to be inserted on the 1 mm long interconnect. Thus, because of a high R for the metal-6 interconnect, a large repeater density is needed to meet the throughput requirement.

The bit rate of 23 Gbps may be achieved on shorter interconnects on metal-6. However, in an actual interconnect system, these shorter interconnects may be routed below metal-6, which could translate into a further increase in R as a result of smaller dimensions, thereby lowering the performance. To reduce R , if the interconnect is routed on metal-10 with a 520 nm pitch and ρ of 2.2 $\mu\text{ohm-cm}$, the required throughput of 23 Gbps is easily achieved by wave-pipelining the interconnect with 2 repeaters per mm. Figure 7.9 shows the repeater density needed to achieve a 23 Gbps throughput if the 1 mm long interconnect is routed on various metal levels between metal-6 and metal-10. It is seen from Figure 7.9 that because of larger interconnect dimensions, ρ decreases for higher metal levels, which reduces R . As a result, the required throughput is achieved by using a smaller repeater density on higher metal levels.

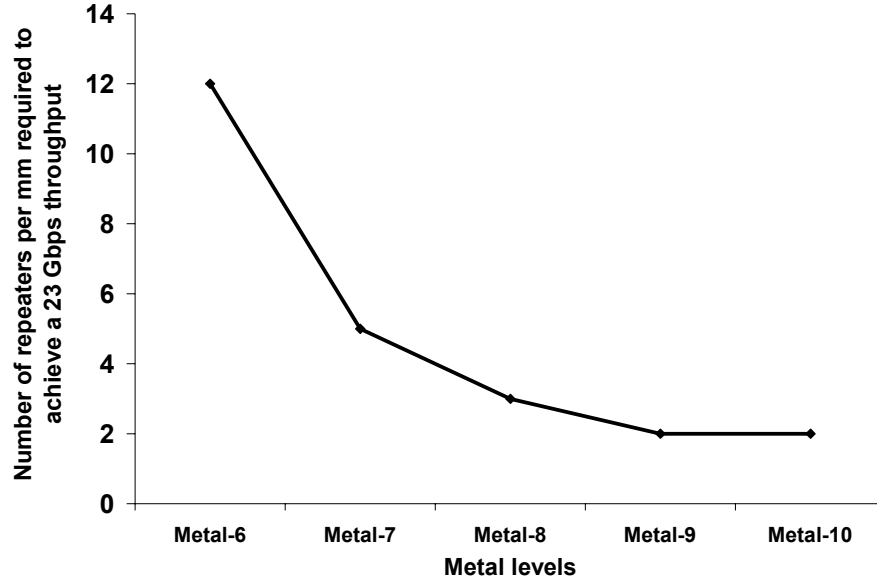


Figure 7.9: Repeater density to achieve a 23 Gbps throughput on different metal levels in a 32 nm node.

7.7.2 Difficulties in achieving low permittivity for interlayer dielectric

Though a relative permittivity of the dielectric (ϵ_r) of less than two is projected for the technology nodes of 32 nm and beyond, the manufacturing solutions to achieve this low ϵ_r are not known [2]. If ϵ_r could not be reduced below its value of 3.1 corresponding to the 90 nm technology, its effect on the throughput is analyzed in this subsection. In general, a high value of ϵ_r results in a higher C , which reduces throughput.

It is intuitive that if C further increases, achieving a 23 Gbps throughput on the 1 mm interconnect on metal-6 in a 32 nm technology generation would require a non-realistic repeater density, and it would therefore be impossible. Therefore, the interconnect is assumed to be routed on metal-10. The required performance is achieved on the wave-pipelined interconnect by inserting three repeaters per mm, even with an ϵ_r of 3.1.

To underline the impact of ϵ_r on throughput, the repeater density required on the 1 mm interconnect on metal-10 is shown in Figure 7.10 for different values of ϵ_r . It is seen that an increase in ϵ_r translates into an increase in C , and a larger repeater density is needed to achieve the same throughput. However, it is important to note that wave-pipelining *achieves the required throughput* even with high values of ϵ_r , unlike latency-centric repeater insertion. This example clearly shows that wave-pipelining can meet the performance requirements on future interconnects, even if some of the projected improvements in certain parameters are not achieved.

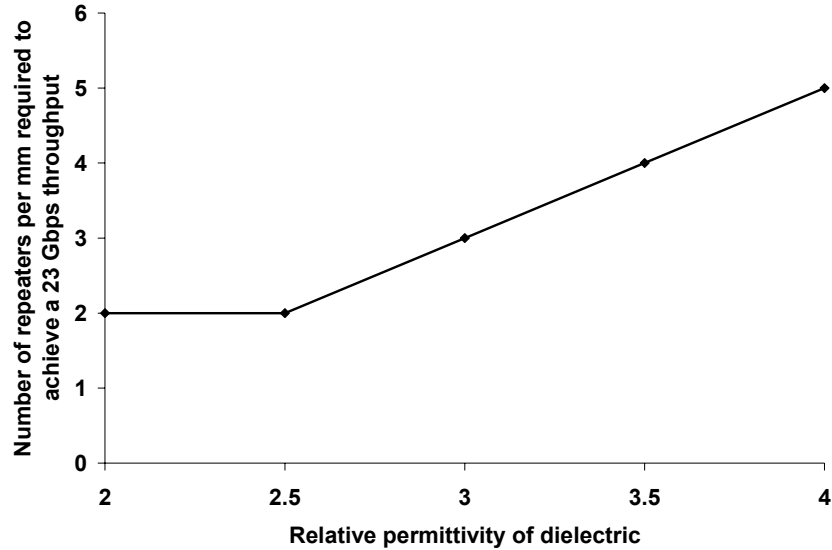


Figure 7.10: Repeater density to achieve a 23 Gbps throughput on metal-10 in a 32 nm node for different values of relative dielectric permittivity.

7.7.3 Insignificant mobility enhancement and large oxide thickness

It is seen in Section 7.2 that the enhancement of mobility and scaling of the gate oxide determine the values of R_t and C_t , respectively. A mobility enhancement factor of two is predicted for the future technology generations, but the manufacturing solutions to

achieve this are not yet known. Similarly, the gate oxide thickness of 0.5 nm is projected for the technology nodes beyond 32 nm, but the manufacturing solutions are unknown [2].

If the mobility enhancement factor does not increase beyond its value of 1.3 (for which manufacturing solutions are known) at the 45 nm node, R_t for the technology nodes beyond 45 nm will increase by more than 50% compared to its values shown in Table 7.2. Similarly, if the oxide thickness does not scale beyond its value of 1 nm (for which manufacturing solutions are known) at the 70 nm node, C_t for the technology nodes beyond 70 nm will be twice of that shown in Table 7.3. This increase in R_t and C_t will not only reduce the throughput of the optimal throughput-centric and latency-centric designs, but it will also affect the *saturation throughput* of wave-pipelined interconnects, which is given by (2.30). Therefore, the limitations imposed by R_t and C_t are more fundamental in nature because they affect the saturation throughput, which sets the upper limit on the maximum bit rate that can be achieved on an interconnect.

7.8 Summary

The analytical throughput model for the wave-pipelined interconnects, which is derived in Chapter 2, is effectively used in this chapter to project the interconnect performance for the future technology generations. The impact of technology scaling on various transistor and interconnect parameters, and subsequently, the interconnect performance, is analyzed in this chapter. The projected values of the throughput for wave-pipelined interconnects are compared to the performance requirements projected by ITRS, and the key issues in enhancing the performance of future interconnect systems are discussed.

Assuming the underlying parameters meet their projected values, wave-pipelining can successfully meet or outperform the ITRS projections, which makes it a very useful technique for the interconnects in future technology generations. Most importantly, this analysis of the future of wave-pipelining also uncovers the immense strength and predictive capability of a simple performance model derived earlier in this research, which highlights the need for such models to characterize important physical and system parameters.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This chapter discusses the possible opportunities related to the refinement and potential extensions to this research. The future work on this thesis primarily includes improving the physical and analytical models, studying the impact of manufacturing and process variations on wave-pipelining, analyzing wave-pipelining for different wiring net models, and extending the system-level analysis of wave-pipelining. A brief discussion of these tasks is presented in this chapter, and the key conclusions of the dissertation are summarized.

8.1 Future work

8.1.1 Future work: improvement of physical and analytical models

The discussion in Section 3.4 suggests that the RC characterization of the interconnect results in a minimal error in the interconnect performance in the region where wave-pipelining is primarily applied. The HSPICE validation of analytical throughput model for RC interconnects and its simple extension for RLC interconnects also support this fact. However, this fact can be further validated by rigorously developing performance models for the RLC interconnects. Deriving a closed-form

expression for the throughput of wave-pipelined RLC interconnects that captures the impact of inductance and number of repeaters on the signal rise time can provide further insights into the behavior of interconnect throughput. Such a model will help gain a better understanding of the boundary between the RC and RLC regimes for the interconnect performance and is therefore desirable.

Better physical models for some of the underlying parameters will also facilitate a more accurate modeling of the interconnect performance. For instance, the transistor resistance R_t has a significant impact on the performance of interconnect circuits. This nonlinear resistance is either modeled by its first-order linear approximation [6] or its value in the saturation regime. Though the evaluation of an intermediate value of R_t is described in Appendix A, it still assumes R_t to be constant throughout the entire operation, which may not be true in real applications. More accurate, voltage-dependent, nonlinear modeling of R_t will enable a more accurate performance analysis of the transistor as well as interconnect circuits. Better models for the interconnect parameters that include certain high-frequency phenomena such as the skin-effect (for R) or modern design trends such as the non-homogeneous dielectric (for C) will also result in a more accurate characterization of the interconnect performance.

8.1.2 Future work: analysis of manufacturing and process variations

The performance variations resulting from the inductive and capacitive coupling for different switching patterns are studied in this research. The impact of power supply fluctuations, which is the largest source of on-chip noise, on the interconnect performance is also included in this research. However, manufacturing and process

variations could also cause variations in the performance. The process-related variations in the dielectric constant, resistivity, threshold voltage, etc., and the manufacturing variations in the dimensions of transistors and interconnects can alter some of the circuit parameters, thereby changing the circuit performance. A careful analysis of the impact of manufacturing and process variations on the throughput of the wave-pipelined interconnect circuits can give a better understanding of the reliability of these interconnect circuits and their tolerance requirements.

By considering the manufacturing and process-related variations at the device level, Monte-Carlo simulations can be performed to study the corresponding variations in the throughput using the analytical model in (2.16) or the latency using the analytical models in [6]. Closed-form expressions can also be derived to establish the correlation between the process-related variations and corresponding performance variations. The impact of manufacturing and process variations on different design optimizations proposed in Chapter 5 can also be studied, and new design optimization rules that minimize the impact of these variations on the circuit performance can be formulated.

8.1.3 Future work: analysis of wave-pipelining for different wiring net models

This research has primarily focused on interconnects that have a single source and a single sink. However, different wiring net models, in which an interconnect is shared between multiple sources and sinks, may exist in actual processor systems. Wave-pipelining can be analyzed for such multiple-source, multiple-sink interconnect networks. An example of such an interconnect is shown in Figure 8.1. Additional timing and control

circuitry, such as tri-state buffers, may be needed to support the wiring net model shown in Figure 8.1. The analysis of wave-pipelining can be extended to the design of control circuitry and calculation of area and power overhead for different wiring net models.

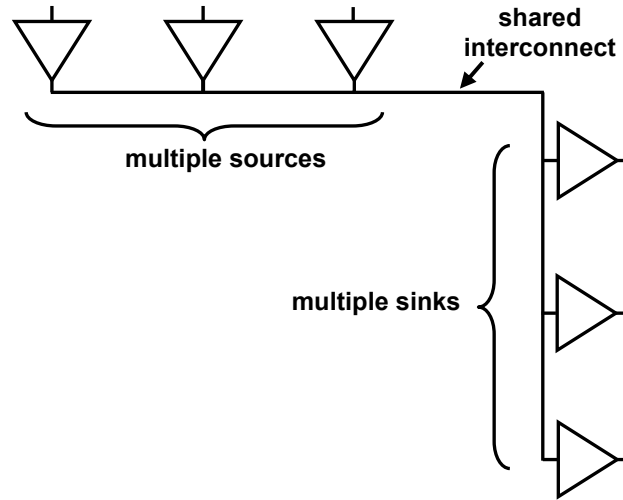


Figure 8.1: A multiple-source, multiple-sink wiring net model.

The analysis of wave-pipelining can also be extended for bidirectional interconnects, which are needed for certain on-chip bus applications. Traditionally, a pair of two unidirectional interconnects is used to construct a bidirectional interconnect. However, special bidirectional repeaters can be constructed so that only a single interconnect can be used as a bidirectional interconnect. The performance-, power-, and area-optimizations for different unidirectional and bidirectional wave-pipelined wiring nets can also be performed using the metrics proposed in Chapter 5.

8.1.4 Future work: extension of system-level analysis

This research has presented the design optimizations for high-performance buses and proposed various timing circuits to integrate wave-pipelined interconnect circuits into

the fully synchronous systems and the systems-on-chip (SoC) using a globally asynchronous locally synchronous (GALS) scheme. However, these analyses can be automated and further extended to apply to the entire interconnect network in a system. Combining the techniques and optimizations in this research with the interconnect distribution models in [51], the holistic performance-power-area analysis for an entire interconnect network can be performed.

For instance, *all* semi-global and global interconnects in an actual processor architecture can be optimized through the simultaneous application of voltage scaling, repeater insertion, and wire sizing to meet a certain performance requirement. This can be achieved through the modification of a multilevel network design simulator, such as MINDS [53], which uses the interconnect distribution model in [51]. The impact of wave-pipelining on the number of metal levels, repeater area, and total interconnect power can be then calculated. The total area and power overhead of the synchronization circuits for wave-pipelined interconnects can also be estimated in this analysis.

8.2 Salient conclusions of dissertation

The primary objective of this research has been to study the circuit- and system-level opportunities of voltage scaling, wire sizing, and repeater insertion in wave-pipelined global interconnect networks that are implemented in deep submicron (DSM) technologies. To meet this objective, this research makes contributions to knowledge in the areas of modeling, design, and simulation of a global interconnect network.

To conclude this dissertation, the main contributions of this research are summarized as follows:

1. *Dramatic improvement over reciprocal latency*

The need for shifting the focus of VLSI interconnect designs from the conventional latency-centric approach to the throughput-centric approach is underlined in this research to enhance the interconnect performance beyond the reciprocal latency. The application of wave-pipelining using repeaters is suggested to achieve this on VLSI global interconnects. It is shown for a global interconnect with a 250 nm x 250 nm square cross-section that the maximum bit rate predicted by the throughput-centric repeater insertion is seven times larger than the maximum reciprocal latency.

2. *Closed-form analytical throughput model*

A simple closed-form analytical expression is derived to calculate the throughput of wave-pipelined RC interconnects. This expression is successfully validated for the 180 nm technology node with HSPICE simulations using RLC interconnects, with an average absolute error of 14%.

3. *Scaling dependence of wave-pipelined throughput*

The analytical throughput model is used to study the impact of interconnect scaling, transistor scaling, and constant field scaling on the interconnect throughput. It is also shown using the analytical throughput model that the maximum saturation throughput that can be achieved on an interconnect circuit is a function of only the technology-dependent parameters. It is shown that the constant field scaling improves the maximum saturation throughput by the scaling factor S .

4. *Advantages of wave-pipelined interconnects over low-loss VLSI transmission lines*

It is shown that wave-pipelining using repeaters can not only enhance the interconnect throughput and signal integrity beyond that obtained on the transmission

line, but it also offers the designer a larger design space, which increases the design flexibility and creates an opportunity to reduce area and power. It is shown that a global interconnect that uses 180 nm repeater circuits results in a 15% reduction in silicon area and a 5% reduction in power compared to a low-loss VLSI transmission line having identical interconnect dimensions, without any loss of throughput performance.

5. *Wave-pipelining for dampening the impact of inductance on performance*

It is shown that the insertion of repeaters makes the effective resistance of an interconnect segment more dominant compared to the characteristic impedance, which dampens the impact of inductance on the interconnect performance. HSPICE simulations show that ignoring interconnect inductance results in a less than 1% error in the throughput of a $1\text{ }\mu\text{m} \times 2.5\text{ }\mu\text{m}$ interconnect in the 180 nm technology when it is wave-pipelined using more than 10 repeaters per cm.

6. *Voltage scaling repeater insertion for high-performance, low-power interconnects*

The simultaneous application of voltage scaling and repeater insertion (VSRI) is proposed to achieve high performance on low-power interconnects. The optimal supply voltage for high-performance, low-power interconnects is shown to be twice of threshold voltage.

7. *Advantages of wave-pipelining over latch insertion and LVDS*

For an identical throughput performance, wave-pipelining is shown to be more area-, power-, and latency-efficient than other high-performance or low-power design techniques such as latch insertion and low-voltage differential signaling (LVDS). It is shown using 180 nm HSPICE simulations that a wave-pipelined metal-5 interconnect

results in a 70% reduction in latency, a 40% reduction in power and an 80% reduction in silicon area compared to the latch-inserted metal-5 interconnect, without any loss of throughput performance. A wave-pipelined overdesign of a global interconnect in the 180 nm technology is shown to completely eliminate the impact of power supply noise on throughput, along with a 40% reduction in power and a 50% reduction in wire area, compared to LVDS.

8. *Application-specific metrics for wave-pipelined design optimizations*

The design optimizations using the simultaneous application of voltage scaling, wire sizing, and repeater insertion are performed to achieve high performance on global interconnect circuits at the expense of minimal power and area. Different design metrics are proposed to optimize interconnect circuits for different applications. For a throughput performance of 1.25 Gbps, the design optimization using the holistic throughput-per-energy-area (TPEA) metric results in a 50% reduction in power and a 90% reduction in wire area compared to the optimal latency-centric design for the 180 nm technology node.

9. *Construction of clock-skew-insensitive receiver circuits*

A new receiver circuit is proposed to aid the communication on wave-pipelined system-on-chip (SoC) global interconnects that connect different clock domains using a globally asynchronous locally synchronous (GALS) scheme. The receiver circuit has a relatively low area and power overhead and it can correctly latch data using a local clock that has a completely random phase w.r.t. the source clock. The complete circuit-level analysis of the wave-pipelined interconnect with this receiver for the 180 nm technology node shows that the interconnect circuit can achieve a high throughput

of 4 Gbps along with clock-skew tolerance of 360° , at the expense of 20 mW of power and $1.8\text{E-}04\text{cm}^2$ of wire area.

10. Impact of technology scaling on wave-pipelining based on ITRS projections

The simple analytical throughput model is exploited to its fullest potential by using it to project the interconnect performance for future technology generations. Wave-pipelining is predicted to be a very effective solution to meet or outperform the ITRS projections for future technology generations. A 1 cm long global interconnect in the 18 nm technology is shown to meet the projected throughput of 53 Gbps using 16 repeaters per cm at the expense of 2.94 mW of power.

Additionally, a variety of research tasks can be performed as an extension to this research. These tasks include improving the physical and analytical models, studying the impact of manufacturing and process variations on wave-pipelining, performing wave-pipelining analysis for different wiring net models, and extending the system-level analysis of wave-pipelining.

APPENDIX A

ESTIMATION OF TRANSISTOR RESISTANCE

The output resistance of a MOSFET, R_t , is a nonlinear function of the supply voltage V_{dd} . Different approximations are used for modeling R_t . The first-order approximation for R_t is given in [6] as

$$R_t \approx \frac{1}{\mu C_{ox} \left(\frac{W}{L} \right) (V_{dd} - |V_t|)} , \quad (\text{A.1})$$

where μ is the electron or hole mobility, C_{ox} is the oxide capacitance, (W/L) is the transistor aspect ratio, and V_t is the threshold voltage. However, (A.1) expresses R_t as a linear function of V_{dd} and is valid only in the linear region of the I-V curve of the MOSFET, before saturation occurs.

Another approximation for R_t estimates its value in the saturation region. In this method, R_t is calculated as

$$R_t = \frac{V_{dd}}{I_{d,sat}} , \quad (\text{A.2})$$

where $I_{d,sat}$ is the saturation drive current that can be obtained from [4].

Figure A.1 shows the I-V curve resulting from HSPICE simulations using level-49 MOSIS transistor models [48] for a 180nm NMOS with a scaling factor of 56. The

NMOS operates from a 2 V supply. The values of R_t obtained from Figure A.1 and their approximations are compared in Table A.1.

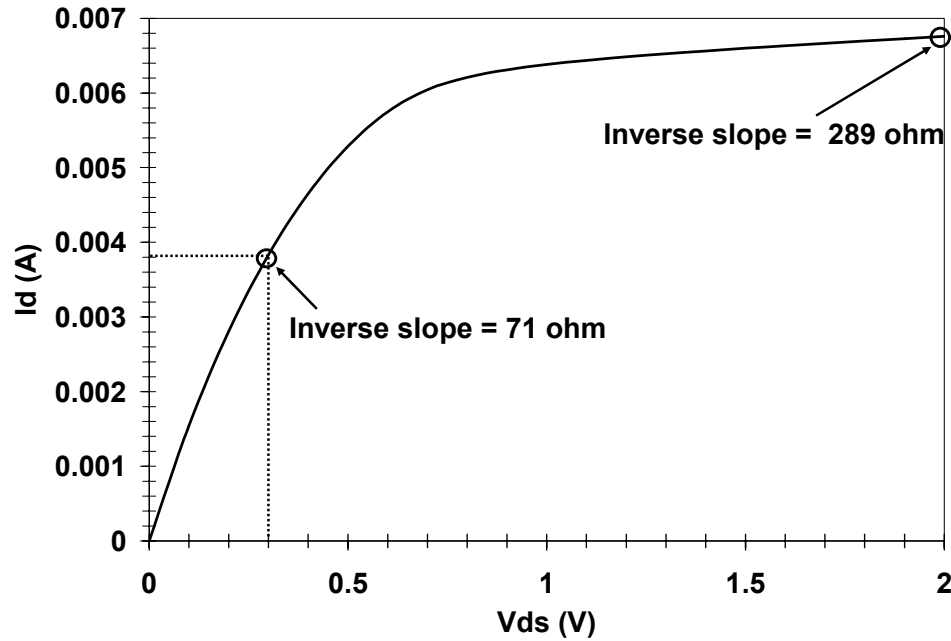


Figure A.1: I-V curve for a 180 nm NMOS.

Table A.1: Values of R_t from I-V curve and approximations.

Region	R_t based on I-V curve in Figure A.1	R_t based on theoretical approximations
Linear	71 ohm	58 ohm (A.1)
Saturation	289 ohm	326 ohm (A.2)

It is seen from Table A.1 that both (A.1) and (A.2) estimate R_t in two extreme regions. Because of the nonlinear nature of the I-V curve, neither of (A.1) or (A.2) completely describes R_t . Moreover, Table A.1 shows that both (A.1) and (A.2) result in an almost 15% error in predicting R_t in the respective regions. Therefore, for a more accurate estimation of R_t , HSPICE-based results should be used. To obtain the typical value of R_t , an average of the two extreme values (71 and 289, based on HSPICE simulations) is considered in this research, which results in an R_t of 180 ohm for the 180 nm transistor having a scaling factor of 56.

APPENDIX B

WAVEFORMS AT DIFFERENT NODES FOR A WAVE-PIPELINED INTERCONNECT

HSPICE results for voltage swings at the input of the interconnect, the output of the first repeated segment, and the output of the interconnect circuit are shown in Figure B.1 for a 1 cm long wave-pipelined interconnect. The interconnect has cross-sectional dimensions of 250 nm x 250 nm, and the interconnect parameters are shown in Table 2.3. It is assumed that 10 repeaters are inserted on the interconnect.

It is seen in Figure B.1 that it is necessary to have a voltage swing that is larger than 90% of the supply voltage at the output of the first repeated segment in order to achieve a 90% voltage swing at the output of the interconnect. This also ensures that the pulsewidth of the input waveform is sufficiently large to avoid any intersymbol interference (ISI) on the interconnect circuit. It is seen in Figure B.1 that all data bits that are transmitted at the input are successfully captured at the output of the interconnect after an approximate delay of 1.5 ns.

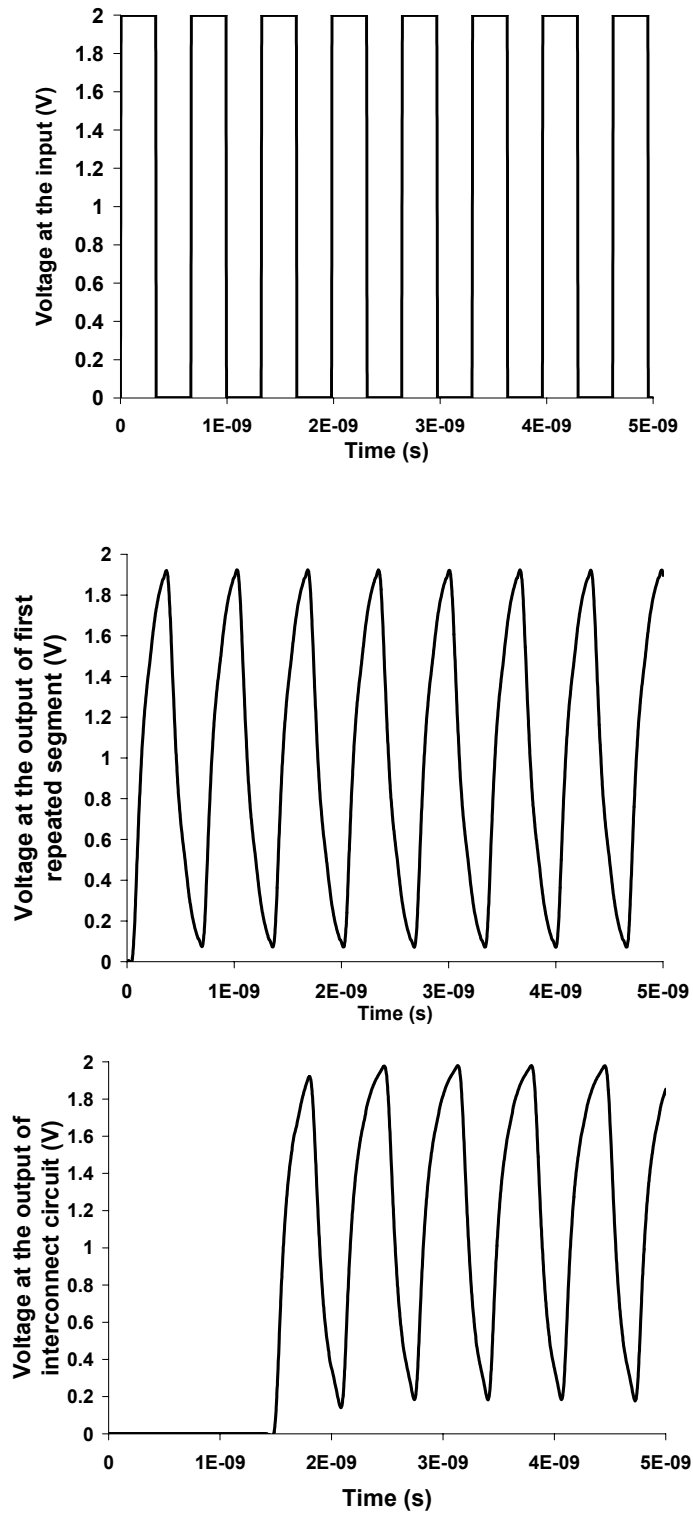


Figure B.1: HSPICE results for voltages at the input, first repeated segment, and output of a wave-pipelined interconnect having a 250 nm x 250 nm cross-section.

APPENDIX C

VALUES OF DESIGN METRICS FOR DIFFERENT DESIGN CHOICES

The interconnect width w is varied from 0.1 μm to 1 μm , the spacing (s)-to-width ratio is varied from 1 to 5, and the interconnect height-to-width aspect ratio (AR) is varied from 1 to 2.5, based on [29]. The dielectric thickness t is assumed to be equal to the metal height h , to meet the constraints on the via aspect ratio [4]. The interconnect is assumed to be placed between two co-planar interconnects and two orthogonal routing planes. Figure 5.3 is redrawn as Figure C.1 to show the interconnect geometry, and the interconnect parasitics for this geometry are extracted using RAPHAEL. The repeater density on the interconnect is varied from 1 to 50 repeaters per cm. The values of different design metrics are shown for all design configurations in this design space, and the optimal values are highlighted. Table C.1 contains the TPBE values, Table C.2 contains the TPEA values, Table C.3 contains the TPA values, and Table C.4 contains the latency values, respectively. The TPBE and TPEA values are shown for a 1 V supply, and the TPA and latency values are shown for a 2 V supply.

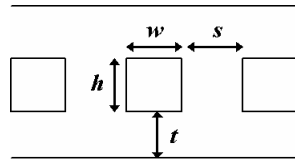


Figure C.1: Interconnect dimensions.

Table C.1: Values of TPBE in Gbps/pJ in design space.

w, s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1	0.005	0.087	0.166	0.216	0.217	0.204	0.190
0.1,0.2	1	0.006	0.099	0.182	0.230	0.227	0.212	0.196
0.1,0.3	1	0.006	0.100	0.184	0.231	0.228	0.213	0.197
0.1,0.4	1	0.006	0.101	0.184	0.232	0.229	0.213	0.197
0.1,0.5	1	0.006	0.101	0.184	0.232	0.229	0.213	0.197
0.2,0.2	1	0.034	0.320	0.458	0.458	0.399	0.343	0.299
0.2,0.4	1	0.042	0.367	0.507	0.490	0.420	0.358	0.309
0.2,0.6	1	0.043	0.373	0.513	0.494	0.423	0.359	0.311
0.2,0.8	1	0.043	0.374	0.514	0.494	0.423	0.360	0.311
0.2,1	1	0.043	0.373	0.513	0.494	0.423	0.359	0.311
0.3,0.3	1	0.081	0.507	0.631	0.564	0.467	0.390	0.332
0.3,0.6	1	0.100	0.585	0.701	0.606	0.493	0.407	0.345
0.3,0.9	1	0.102	0.594	0.709	0.611	0.496	0.409	0.346
0.3,1.2	1	0.103	0.595	0.710	0.611	0.496	0.409	0.346
0.3,1.5	1	0.103	0.595	0.710	0.611	0.496	0.409	0.346
0.4,0.4	1	0.131	0.626	0.723	0.614	0.497	0.409	0.346
0.4,0.8	1	0.161	0.724	0.805	0.660	0.525	0.427	0.359
0.4,1.2	1	0.165	0.736	0.815	0.665	0.528	0.429	0.361
0.4,1.6	1	0.166	0.737	0.816	0.665	0.528	0.430	0.361
0.4,2	1	0.166	0.737	0.816	0.665	0.528	0.430	0.361
0.5,0.5	1	0.176	0.702	0.775	0.640	0.512	0.419	0.353
0.5,1	1	0.216	0.812	0.864	0.687	0.541	0.438	0.366
0.5,1.5	1	0.221	0.825	0.874	0.693	0.544	0.440	0.368
0.5,2	1	0.222	0.827	0.875	0.694	0.544	0.440	0.368
0.5,2.5	1	0.222	0.826	0.875	0.693	0.544	0.440	0.368
0.6,0.6	1	0.214	0.753	0.809	0.656	0.521	0.425	0.357
0.6,1.2	1	0.263	0.871	0.901	0.705	0.550	0.444	0.370
0.6,1.8	1	0.270	0.888	0.914	0.711	0.554	0.446	0.372
0.6,2.4	1	0.271	0.890	0.916	0.712	0.555	0.446	0.372
0.6,3	1	0.271	0.890	0.916	0.712	0.555	0.446	0.372
0.7,0.7	1	0.243	0.785	0.829	0.665	0.526	0.428	0.359
0.7,1.4	1	0.300	0.911	0.926	0.716	0.556	0.448	0.373
0.7,2.1	1	0.308	0.927	0.938	0.722	0.560	0.450	0.375
0.7,2.8	1	0.309	0.929	0.939	0.722	0.560	0.450	0.375
0.7,3.5	1	0.309	0.929	0.939	0.722	0.560	0.450	0.375
0.8,0.8	1	0.267	0.807	0.842	0.671	0.530	0.430	0.361
0.8,1.6	1	0.330	0.938	0.942	0.723	0.560	0.450	0.375
0.8,2.4	1	0.338	0.954	0.953	0.728	0.564	0.452	0.376
0.8,3.2	1	0.338	0.955	0.954	0.729	0.564	0.452	0.376
0.8,4	1	0.339	0.956	0.954	0.729	0.564	0.452	0.376
0.9,0.9	1	0.286	0.824	0.852	0.675	0.532	0.432	0.362
0.9,1.8	1	0.353	0.957	0.952	0.727	0.563	0.451	0.376
0.9,2.7	1	0.362	0.974	0.964	0.733	0.566	0.454	0.377
0.9,3.6	1	0.363	0.976	0.966	0.734	0.567	0.454	0.377
0.9,4.5	1	0.363	0.975	0.966	0.734	0.567	0.454	0.377
1,1	1	0.301	0.835	0.859	0.678	0.534	0.433	0.363
1,2	1	0.371	0.968	0.958	0.729	0.564	0.452	0.376
1,3	1	0.381	0.987	0.972	0.736	0.568	0.455	0.378
1,4	1	0.382	0.989	0.973	0.737	0.568	0.455	0.378
1,5	1	0.380	0.986	0.971	0.736	0.568	0.455	0.378

[continued] Table C.1: Values of TPBE in Gbps/pJ in design space.

w, s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1.5	0.008	0.129	0.229	0.278	0.268	0.246	0.224
0.1,0.2	1.5	0.012	0.170	0.279	0.318	0.297	0.267	0.240
0.1,0.3	1.5	0.013	0.180	0.291	0.326	0.303	0.271	0.244
0.1,0.4	1.5	0.014	0.182	0.294	0.328	0.304	0.272	0.244
0.1,0.5	1.5	0.014	0.183	0.295	0.329	0.305	0.273	0.245
0.2,0.2	1.5	0.050	0.389	0.525	0.501	0.428	0.363	0.313
0.2,0.4	1.5	0.076	0.520	0.650	0.578	0.477	0.396	0.337
0.2,0.6	1.5	0.083	0.553	0.680	0.596	0.487	0.403	0.342
0.2,0.8	1.5	0.085	0.561	0.687	0.600	0.490	0.405	0.343
0.2,1	1.5	0.086	0.563	0.689	0.601	0.490	0.405	0.343
0.3,0.3	1.5	0.106	0.559	0.668	0.583	0.478	0.397	0.337
0.3,0.6	1.5	0.162	0.752	0.832	0.675	0.534	0.434	0.364
0.3,0.9	1.5	0.178	0.800	0.870	0.695	0.546	0.441	0.369
0.3,1.2	1.5	0.182	0.812	0.879	0.700	0.549	0.443	0.370
0.3,1.5	1.5	0.183	0.815	0.882	0.701	0.550	0.444	0.371
0.4,0.4	1.5	0.158	0.655	0.737	0.618	0.499	0.410	0.347
0.4,0.8	1.5	0.242	0.884	0.921	0.717	0.558	0.449	0.374
0.4,1.2	1.5	0.266	0.942	0.964	0.739	0.571	0.457	0.380
0.4,1.6	1.5	0.273	0.961	0.978	0.746	0.574	0.459	0.381
0.4,2	1.5	0.275	0.964	0.980	0.747	0.575	0.460	0.382
0.5,0.5	1.5	0.200	0.711	0.774	0.636	0.509	0.417	0.352
0.5,1	1.5	0.306	0.961	0.968	0.738	0.570	0.456	0.379
0.5,1.5	1.5	0.338	1.029	1.016	0.762	0.583	0.465	0.385
0.5,2	1.5	0.346	1.045	1.028	0.768	0.586	0.467	0.386
0.5,2.5	1.5	0.348	1.049	1.031	0.769	0.587	0.467	0.387
0.6,0.6	1.5	0.232	0.746	0.796	0.647	0.515	0.421	0.354
0.6,1.2	1.5	0.355	1.009	0.996	0.750	0.576	0.460	0.382
0.6,1.8	1.5	0.392	1.080	1.046	0.774	0.590	0.469	0.388
0.6,2.4	1.5	0.401	1.096	1.057	0.780	0.593	0.471	0.389
0.6,3	1.5	0.404	1.102	1.061	0.782	0.594	0.472	0.390
0.7,0.7	1.5	0.255	0.768	0.809	0.653	0.519	0.423	0.356
0.7,1.4	1.5	0.394	1.042	1.015	0.758	0.581	0.463	0.384
0.7,2.1	1.5	0.432	1.110	1.062	0.781	0.594	0.471	0.389
0.7,2.8	1.5	0.443	1.129	1.075	0.787	0.597	0.473	0.391
0.7,3.5	1.5	0.445	1.134	1.078	0.789	0.598	0.474	0.391
0.8,0.8	1.5	0.274	0.783	0.818	0.657	0.521	0.424	0.357
0.8,1.6	1.5	0.421	1.062	1.025	0.763	0.583	0.464	0.385
0.8,2.4	1.5	0.463	1.132	1.074	0.786	0.596	0.473	0.390
0.8,3.2	1.5	0.475	1.153	1.088	0.793	0.600	0.475	0.392
0.8,4	1.5	0.478	1.157	1.091	0.794	0.601	0.476	0.392
0.9,0.9	1.5	0.288	0.795	0.826	0.660	0.523	0.426	0.358
0.9,1.8	1.5	0.445	1.080	1.036	0.767	0.585	0.466	0.386
0.9,2.7	1.5	0.490	1.154	1.086	0.791	0.599	0.475	0.392
0.9,3.6	1.5	0.502	1.172	1.099	0.797	0.602	0.477	0.393
0.9,4.5	1.5	0.505	1.176	1.102	0.798	0.603	0.477	0.393
1,1	1.5	0.299	0.803	0.830	0.662	0.524	0.426	0.358
1,2	1.5	0.462	1.091	1.042	0.770	0.587	0.467	0.386
1,3	1.5	0.509	1.166	1.093	0.794	0.600	0.475	0.392
1,4	1.5	0.521	1.184	1.105	0.799	0.604	0.477	0.394
1,5	1.5	0.524	1.188	1.108	0.801	0.604	0.478	0.394

[continued] Table C.1: Values of TPBE in Gbps/pJ in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2	0.010	0.151	0.260	0.308	0.292	0.265	0.240
0.1,0.2	2	0.018	0.224	0.348	0.374	0.339	0.299	0.265
0.1,0.3	2	0.022	0.250	0.376	0.394	0.353	0.309	0.272
0.1,0.4	2	0.023	0.259	0.386	0.401	0.357	0.312	0.275
0.1,0.5	2	0.024	0.262	0.389	0.403	0.359	0.313	0.275
0.2,0.2	2	0.056	0.396	0.526	0.499	0.426	0.361	0.312
0.2,0.4	2	0.100	0.598	0.716	0.615	0.499	0.411	0.347
0.2,0.6	2	0.120	0.673	0.779	0.650	0.520	0.424	0.357
0.2,0.8	2	0.127	0.699	0.801	0.661	0.527	0.429	0.360
0.2,1	2	0.129	0.708	0.808	0.665	0.529	0.431	0.361
0.3,0.3	2	0.109	0.533	0.637	0.562	0.464	0.387	0.331
0.3,0.6	2	0.198	0.815	0.875	0.695	0.546	0.441	0.369
0.3,0.9	2	0.237	0.919	0.954	0.736	0.569	0.456	0.379
0.3,1.2	2	0.252	0.956	0.981	0.749	0.577	0.461	0.383
0.3,1.5	2	0.257	0.970	0.991	0.754	0.580	0.463	0.384
0.4,0.4	2	0.154	0.604	0.687	0.587	0.480	0.397	0.338
0.4,0.8	2	0.279	0.928	0.946	0.728	0.564	0.453	0.377
0.4,1.2	2	0.333	1.044	1.031	0.770	0.588	0.468	0.387
0.4,1.6	2	0.355	1.090	1.063	0.785	0.597	0.473	0.391
0.4,2	2	0.363	1.106	1.074	0.790	0.599	0.475	0.392
0.5,0.5	2	0.186	0.643	0.712	0.600	0.487	0.402	0.341
0.5,1	2	0.338	0.987	0.981	0.743	0.572	0.458	0.380
0.5,1.5	2	0.406	1.116	1.072	0.787	0.597	0.474	0.391
0.5,2	2	0.431	1.161	1.103	0.802	0.605	0.479	0.395
0.5,2.5	2	0.440	1.177	1.113	0.807	0.608	0.480	0.396
0.6,0.6	2	0.211	0.668	0.728	0.607	0.491	0.405	0.343
0.6,1.2	2	0.383	1.026	1.004	0.753	0.578	0.461	0.382
0.6,1.8	2	0.459	1.159	1.096	0.797	0.603	0.477	0.393
0.6,2.4	2	0.487	1.207	1.128	0.812	0.611	0.482	0.397
0.6,3	2	0.501	1.228	1.142	0.818	0.614	0.484	0.398
0.7,0.7	2	0.229	0.684	0.738	0.612	0.494	0.406	0.344
0.7,1.4	2	0.417	1.053	1.019	0.760	0.581	0.463	0.384
0.7,2.1	2	0.501	1.192	1.114	0.805	0.607	0.479	0.395
0.7,2.8	2	0.533	1.243	1.148	0.820	0.615	0.485	0.399
0.7,3.5	2	0.545	1.260	1.159	0.825	0.618	0.486	0.400
0.8,0.8	2	0.242	0.694	0.745	0.615	0.495	0.407	0.345
0.8,1.6	2	0.441	1.069	1.028	0.763	0.583	0.464	0.385
0.8,2.4	2	0.530	1.211	1.124	0.809	0.609	0.481	0.396
0.8,3.2	2	0.564	1.262	1.158	0.824	0.617	0.486	0.399
0.8,4	2	0.575	1.278	1.168	0.829	0.620	0.488	0.401
0.9,0.9	2	0.251	0.701	0.749	0.617	0.497	0.408	0.345
0.9,1.8	2	0.459	1.081	1.034	0.766	0.585	0.465	0.385
0.9,2.7	2	0.552	1.224	1.131	0.812	0.610	0.482	0.396
0.9,3.6	2	0.587	1.276	1.165	0.827	0.619	0.487	0.400
0.9,4.5	2	0.599	1.292	1.175	0.832	0.621	0.489	0.401
1,1	2	0.259	0.706	0.752	0.618	0.497	0.409	0.346
1,2	2	0.473	1.089	1.038	0.768	0.586	0.466	0.386
1,3	2	0.568	1.233	1.136	0.813	0.611	0.482	0.397
1,4	2	0.605	1.286	1.170	0.829	0.620	0.488	0.400
1,5	2	0.617	1.303	1.181	0.834	0.623	0.489	0.402

[continued] Table C.1: Values of TPBE in Gbps/pJ in design space.

w, s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2.5	0.012	0.159	0.272	0.319	0.301	0.272	0.245
0.1,0.2	2.5	0.023	0.260	0.391	0.409	0.365	0.318	0.280
0.1,0.3	2.5	0.030	0.305	0.439	0.441	0.386	0.333	0.291
0.1,0.4	2.5	0.033	0.325	0.460	0.454	0.395	0.339	0.295
0.1,0.5	2.5	0.034	0.334	0.468	0.459	0.398	0.341	0.297
0.2,0.2	2.5	0.056	0.376	0.500	0.480	0.412	0.352	0.305
0.2,0.4	2.5	0.114	0.626	0.735	0.624	0.504	0.414	0.350
0.2,0.6	2.5	0.147	0.743	0.832	0.677	0.536	0.435	0.364
0.2,0.8	2.5	0.163	0.794	0.872	0.698	0.548	0.443	0.370
0.2,1	2.5	0.170	0.816	0.889	0.707	0.553	0.446	0.372
0.3,0.3	2.5	0.103	0.484	0.586	0.528	0.442	0.372	0.320
0.3,0.6	2.5	0.211	0.815	0.869	0.691	0.543	0.439	0.367
0.3,0.9	2.5	0.273	0.970	0.986	0.750	0.577	0.461	0.383
0.3,1.2	2.5	0.301	1.036	1.033	0.773	0.590	0.469	0.388
0.3,1.5	2.5	0.317	1.070	1.057	0.784	0.597	0.473	0.391
0.4,0.4	2.5	0.139	0.536	0.623	0.547	0.454	0.380	0.325
0.4,0.8	2.5	0.284	0.905	0.925	0.716	0.557	0.448	0.373
0.4,1.2	2.5	0.368	1.079	1.050	0.778	0.592	0.471	0.389
0.4,1.6	2.5	0.410	1.159	1.105	0.804	0.607	0.480	0.395
0.4,2	2.5	0.427	1.190	1.126	0.814	0.612	0.483	0.397
0.5,0.5	2.5	0.165	0.565	0.641	0.557	0.459	0.383	0.327
0.5,1	2.5	0.338	0.957	0.956	0.730	0.564	0.453	0.377
0.5,1.5	2.5	0.439	1.143	1.087	0.793	0.601	0.476	0.392
0.5,2	2.5	0.488	1.226	1.142	0.819	0.615	0.485	0.399
0.5,2.5	2.5	0.510	1.263	1.167	0.830	0.621	0.488	0.401
0.6,0.6	2.5	0.183	0.583	0.653	0.562	0.463	0.385	0.329
0.6,1.2	2.5	0.375	0.987	0.973	0.737	0.568	0.455	0.378
0.6,1.8	2.5	0.488	1.181	1.107	0.802	0.605	0.478	0.394
0.6,2.4	2.5	0.542	1.265	1.163	0.828	0.619	0.487	0.400
0.6,3	2.5	0.567	1.303	1.188	0.839	0.625	0.491	0.403
0.7,0.7	2.5	0.195	0.593	0.660	0.566	0.465	0.387	0.330
0.7,1.4	2.5	0.403	1.008	0.985	0.742	0.571	0.457	0.379
0.7,2.1	2.5	0.522	1.202	1.119	0.806	0.607	0.480	0.395
0.7,2.8	2.5	0.580	1.290	1.176	0.833	0.622	0.489	0.401
0.7,3.5	2.5	0.608	1.330	1.202	0.844	0.628	0.493	0.404
0.8,0.8	2.5	0.205	0.601	0.664	0.568	0.466	0.387	0.330
0.8,1.6	2.5	0.422	1.020	0.991	0.745	0.573	0.458	0.380
0.8,2.4	2.5	0.547	1.218	1.127	0.810	0.609	0.481	0.396
0.8,3.2	2.5	0.608	1.307	1.185	0.836	0.624	0.490	0.402
0.8,4	2.5	0.637	1.346	1.210	0.847	0.630	0.494	0.405
0.9,0.9	2.5	0.211	0.605	0.667	0.569	0.467	0.388	0.331
0.9,1.8	2.5	0.436	1.029	0.996	0.747	0.574	0.458	0.381
0.9,2.7	2.5	0.565	1.228	1.132	0.812	0.610	0.482	0.396
0.9,3.6	2.5	0.629	1.318	1.190	0.838	0.625	0.491	0.403
0.9,4.5	2.5	0.657	1.357	1.215	0.849	0.631	0.494	0.405
1,1	2.5	0.216	0.609	0.669	0.570	0.467	0.388	0.331
1,2	2.5	0.447	1.035	1.000	0.748	0.575	0.459	0.381
1,3	2.5	0.579	1.236	1.136	0.813	0.611	0.482	0.397
1,4	2.5	0.644	1.326	1.194	0.840	0.626	0.491	0.403
1,5	2.5	0.673	1.365	1.219	0.851	0.632	0.495	0.405

Table C.2: Values of TPEA in Tbps/pJ-cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1	0.216	3.546	5.731	5.710	4.641	3.675	2.949
0.1,0.2	1	0.179	2.874	4.686	4.809	4.009	3.238	2.638
0.1,0.3	1	0.138	2.260	3.766	4.004	3.425	2.821	2.334
0.1,0.4	1	0.111	1.849	3.132	3.417	2.981	2.493	2.088
0.1,0.5	1	0.093	1.562	2.677	2.978	2.637	2.232	1.888
0.2,0.2	1	0.831	7.192	9.364	7.924	5.985	4.538	3.536
0.2,0.4	1	0.683	5.696	7.360	6.303	4.849	3.743	2.962
0.2,0.6	1	0.527	4.418	5.771	5.053	3.964	3.110	2.496
0.2,0.8	1	0.423	3.577	4.717	4.198	3.340	2.652	2.151
0.2,1	1	0.353	3.001	3.983	3.588	2.884	2.311	1.889
0.3,0.3	1	1.338	7.861	9.152	7.255	5.392	4.077	3.183
0.3,0.6	1	1.100	6.190	7.091	5.622	4.227	3.241	2.565
0.3,0.9	1	0.847	4.773	5.504	4.432	3.382	2.628	2.105
0.3,1.2	1	0.680	3.852	4.470	3.642	2.809	2.204	1.781
0.3,1.5	1	0.567	3.226	3.761	3.090	2.402	1.898	1.543
0.4,0.4	1	1.622	7.415	8.131	6.274	4.657	3.539	2.782
0.4,0.8	1	1.335	5.821	6.248	4.787	3.577	2.747	2.184
0.4,1.2	1	1.026	4.476	4.825	3.739	2.827	2.196	1.764
0.4,1.6	1	0.824	3.607	3.906	3.055	2.330	1.824	1.476
0.4,2	1	0.687	3.016	3.278	2.580	1.980	1.559	1.268
0.5,0.5	1	1.743	6.718	7.119	5.430	4.041	3.089	2.444
0.5,1	1	1.432	5.257	5.438	4.098	3.060	2.358	1.883
0.5,1.5	1	1.101	4.037	4.186	3.182	2.399	1.866	1.504
0.5,2	1	0.884	3.249	3.382	2.590	1.967	1.540	1.249
0.5,2.5	1	0.737	2.715	2.834	2.182	1.666	1.311	1.068
0.6,0.6	1	1.770	6.054	6.279	4.760	3.554	2.731	2.172
0.6,1.2	1	1.452	4.722	4.772	3.563	2.662	2.058	1.650
0.6,1.8	1	1.121	3.633	3.674	2.760	2.078	1.619	1.308
0.6,2.4	1	0.900	2.923	2.964	2.241	1.698	1.330	1.081
0.6,3	1	0.750	2.441	2.482	1.885	1.434	1.129	0.921
0.7,0.7	1	1.727	5.432	5.565	4.213	3.158	2.438	1.948
0.7,1.4	1	1.425	4.250	4.230	3.142	2.351	1.822	1.466
0.7,2.1	1	1.096	3.259	3.245	2.423	1.826	1.425	1.154
0.7,2.8	1	0.880	2.622	2.617	1.964	1.488	1.167	0.950
0.7,3.5	1	0.734	2.189	2.189	1.650	1.254	0.988	0.807
0.8,0.8	1	1.661	4.910	4.988	3.774	2.838	2.200	1.765
0.8,1.6	1	1.370	3.838	3.783	2.803	2.101	1.633	1.317
0.8,2.4	1	1.054	2.941	2.899	2.157	1.626	1.271	1.032
0.8,3.2	1	0.844	2.361	2.333	1.744	1.321	1.038	0.847
0.8,4	1	0.705	1.973	1.952	1.465	1.113	0.877	0.718
0.9,0.9	1	1.582	4.466	4.511	3.414	2.575	2.003	1.613
0.9,1.8	1	1.305	3.488	3.415	2.527	1.897	1.477	1.195
0.9,2.7	1	1.004	2.672	2.615	1.941	1.464	1.147	0.933
0.9,3.6	1	0.806	2.147	2.105	1.569	1.189	0.935	0.763
0.9,4.5	1	0.671	1.791	1.759	1.316	1.000	0.789	0.646
1,1	1	1.499	4.086	4.111	3.114	2.355	1.837	1.484
1,2	1	1.232	3.180	3.102	2.295	1.726	1.348	1.092
1,3	1	0.951	2.441	2.377	1.763	1.331	1.044	0.850
1,4	1	0.763	1.961	1.913	1.424	1.079	0.850	0.695
1,5	1	0.633	1.630	1.594	1.191	0.906	0.715	0.586

[continued] Table C.2: Values of TPEA in Tbps/pJ-cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1.5	0.382	5.260	7.913	7.358	5.746	4.426	3.479
0.1,0.2	1.5	0.391	4.922	7.183	6.645	5.235	4.071	3.227
0.1,0.3	1.5	0.324	4.044	5.956	5.645	4.541	3.590	2.884
0.1,0.4	1.5	0.266	3.349	4.995	4.844	3.968	3.183	2.587
0.1,0.5	1.5	0.224	2.837	4.278	4.226	3.513	2.852	2.341
0.2,0.2	1.5	1.218	8.756	10.740	8.674	6.412	4.800	3.708
0.2,0.4	1.5	1.246	8.066	9.441	7.436	5.498	4.143	3.226
0.2,0.6	1.5	1.030	6.549	7.651	6.091	4.566	3.487	2.748
0.2,0.8	1.5	0.844	5.370	6.310	5.091	3.864	2.984	2.375
0.2,1	1.5	0.708	4.522	5.343	4.358	3.341	2.603	2.088
0.3,0.3	1.5	1.743	8.668	9.692	7.494	5.517	4.151	3.231
0.3,0.6	1.5	1.784	7.958	8.412	6.264	4.579	3.454	2.704
0.3,0.9	1.5	1.471	6.426	6.751	5.046	3.723	2.837	2.244
0.3,1.2	1.5	1.205	5.255	5.535	4.172	3.107	2.388	1.904
0.3,1.5	1.5	1.010	4.416	4.668	3.546	2.659	2.058	1.651
0.4,0.4	1.5	1.953	7.757	8.292	6.324	4.679	3.551	2.788
0.4,0.8	1.5	2.001	7.106	7.143	5.205	3.805	2.885	2.274
0.4,1.2	1.5	1.651	5.731	5.708	4.156	3.057	2.336	1.857
0.4,1.6	1.5	1.361	4.699	4.680	3.424	2.534	1.950	1.560
0.4,2	1.5	1.141	3.943	3.937	2.897	2.157	1.668	1.342
0.5,0.5	1.5	1.978	6.807	7.108	5.401	4.021	3.076	2.434
0.5,1	1.5	2.027	6.225	6.092	4.400	3.225	2.458	1.949
0.5,1.5	1.5	1.682	5.032	4.866	3.499	2.573	1.973	1.575
0.5,2	1.5	1.379	4.107	3.971	2.867	2.120	1.635	1.312
0.5,2.5	1.5	1.156	3.445	3.337	2.420	1.798	1.393	1.123
0.6,0.6	1.5	1.915	5.993	6.176	4.692	3.513	2.704	2.154
0.6,1.2	1.5	1.964	5.473	5.273	3.794	2.789	2.135	1.701
0.6,1.8	1.5	1.628	4.417	4.201	3.004	2.212	1.702	1.363
0.6,2.4	1.5	1.333	3.600	3.422	2.454	1.815	1.403	1.130
0.6,3	1.5	1.120	3.024	2.876	2.069	1.537	1.192	0.963
0.7,0.7	1.5	1.813	5.314	5.435	4.136	3.112	2.409	1.928
0.7,1.4	1.5	1.867	4.861	4.637	3.330	2.454	1.885	1.508
0.7,2.1	1.5	1.537	3.902	3.676	2.623	1.935	1.493	1.200
0.7,2.8	1.5	1.262	3.187	2.996	2.141	1.585	1.228	0.991
0.7,3.5	1.5	1.058	2.671	2.514	1.802	1.339	1.040	0.842
0.8,0.8	1.5	1.702	4.761	4.845	3.694	2.791	2.170	1.745
0.8,1.6	1.5	1.747	4.343	4.119	2.958	2.186	1.685	1.352
0.8,2.4	1.5	1.443	3.491	3.266	2.327	1.720	1.330	1.071
0.8,3.2	1.5	1.185	2.850	2.661	1.897	1.406	1.091	0.882
0.8,4	1.5	0.994	2.389	2.232	1.595	1.186	0.922	0.748
0.9,0.9	1.5	1.595	4.309	4.371	3.337	2.530	1.974	1.593
0.9,1.8	1.5	1.642	3.934	3.713	2.666	1.973	1.525	1.227
0.9,2.7	1.5	1.359	3.166	2.945	2.095	1.549	1.200	0.968
0.9,3.6	1.5	1.113	2.579	2.394	1.704	1.264	0.981	0.795
0.9,4.5	1.5	0.933	2.161	2.007	1.431	1.064	0.829	0.673
1,1	1.5	1.491	3.926	3.974	3.040	2.312	1.809	1.465
1,2	1.5	1.536	3.585	3.373	2.422	1.796	1.391	1.121
1,3	1.5	1.270	2.882	2.672	1.900	1.407	1.091	0.882
1,4	1.5	1.040	2.347	2.171	1.544	1.146	0.891	0.723
1,5	1.5	0.872	1.966	1.819	1.296	0.964	0.752	0.611

[continued] Table. C.2: Values of TPEA in Tbps/pJ-cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2	0.491	6.160	9.004	8.151	6.265	4.772	3.718
0.1,0.2	2	0.594	6.490	8.943	7.833	5.987	4.564	3.565
0.1,0.3	2	0.534	5.618	7.695	6.818	5.289	4.084	3.226
0.1,0.4	2	0.455	4.758	6.556	5.910	4.657	3.644	2.909
0.1,0.5	2	0.388	4.070	5.653	5.179	4.137	3.273	2.638
0.2,0.2	2	1.364	8.910	10.756	8.643	6.388	4.783	3.696
0.2,0.4	2	1.646	9.282	10.388	7.905	5.754	4.297	3.327
0.2,0.6	2	1.479	7.966	8.762	6.646	4.873	3.673	2.870
0.2,0.8	2	1.258	6.693	7.352	5.616	4.158	3.164	2.495
0.2,1	2	1.071	5.691	6.269	4.830	3.607	2.767	2.198
0.3,0.3	2	1.795	8.275	9.246	7.223	5.358	4.052	3.165
0.3,0.6	2	2.180	8.628	8.844	6.452	4.678	3.513	2.742
0.3,0.9	2	1.960	7.386	7.402	5.341	3.881	2.932	2.306
0.3,1.2	2	1.668	6.192	6.177	4.467	3.266	2.485	1.967
0.3,1.5	2	1.422	5.259	5.248	3.814	2.805	2.147	1.710
0.4,0.4	2	1.899	7.158	7.730	6.008	4.495	3.436	2.712
0.4,0.8	2	2.312	7.455	7.343	5.286	3.847	2.910	2.290
0.4,1.2	2	2.069	6.350	6.104	4.330	3.149	2.392	1.893
0.4,1.6	2	1.769	5.334	5.090	3.606	2.632	2.009	1.599
0.4,2	2	1.507	4.524	4.315	3.066	2.248	1.724	1.378
0.5,0.5	2	1.848	6.159	6.543	5.092	3.842	2.963	2.359
0.5,1	2	2.243	6.389	6.174	4.430	3.240	2.467	1.955
0.5,1.5	2	2.019	5.456	5.131	3.614	2.635	2.010	1.600
0.5,2	2	1.717	4.563	4.259	2.994	2.188	1.676	1.340
0.5,2.5	2	1.461	3.866	3.605	2.538	1.862	1.432	1.149
0.6,0.6	2	1.744	5.368	5.650	4.408	3.348	2.601	2.084
0.6,1.2	2	2.119	5.563	5.313	3.807	2.795	2.139	1.704
0.6,1.8	2	1.905	4.742	4.403	3.092	2.260	1.730	1.382
0.6,2.4	2	1.620	3.963	3.650	2.555	1.870	1.436	1.152
0.6,3	2	1.388	3.371	3.096	2.167	1.589	1.224	0.985
0.7,0.7	2	1.624	4.736	4.959	3.880	2.963	2.315	1.865
0.7,1.4	2	1.979	4.911	4.656	3.335	2.456	1.887	1.509
0.7,2.1	2	1.783	4.190	3.856	2.702	1.978	1.519	1.217
0.7,2.8	2	1.520	3.506	3.198	2.230	1.633	1.257	1.010
0.7,3.5	2	1.294	2.968	2.702	1.885	1.384	1.068	0.861
0.8,0.8	2	1.502	4.222	4.408	3.459	2.654	2.083	1.686
0.8,1.6	2	1.832	4.375	4.130	2.961	2.187	1.686	1.352
0.8,2.4	2	1.652	3.732	3.418	2.394	1.756	1.352	1.086
0.8,3.2	2	1.408	3.121	2.832	1.973	1.447	1.116	0.899
0.8,4	2	1.196	2.638	2.389	1.665	1.224	0.946	0.764
0.9,0.9	2	1.390	3.802	3.964	3.119	2.403	1.893	1.538
0.9,1.8	2	1.695	3.938	3.707	2.661	1.970	1.523	1.225
0.9,2.7	2	1.530	3.360	3.067	2.148	1.578	1.217	0.980
0.9,3.6	2	1.302	2.807	2.538	1.768	1.298	1.003	0.809
0.9,4.5	2	1.107	2.373	2.141	1.491	1.097	0.849	0.686
1,1	2	1.288	3.455	3.598	2.839	2.194	1.734	1.413
1,2	2	1.572	3.577	3.362	2.416	1.792	1.388	1.120
1,3	2	1.418	3.050	2.778	1.947	1.433	1.107	0.893
1,4	2	1.208	2.549	2.299	1.601	1.177	0.910	0.736
1,5	2	1.027	2.155	1.939	1.350	0.993	0.770	0.623

[continued] Table C.2: Values of TPEA in Tbps/pJ-cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2.5	0.554	6.517	9.407	8.442	6.456	4.899	3.807
0.1,0.2	2.5	0.753	7.535	10.063	8.557	6.433	4.850	3.758
0.1,0.3	2.5	0.730	6.866	8.989	7.635	5.790	4.407	3.444
0.1,0.4	2.5	0.650	5.974	7.805	6.700	5.146	3.960	3.125
0.1,0.5	2.5	0.566	5.180	6.794	5.907	4.592	3.570	2.842
0.2,0.2	2.5	1.374	8.468	10.226	8.306	6.187	4.658	3.613
0.2,0.4	2.5	1.870	9.717	10.670	8.028	5.818	4.335	3.351
0.2,0.6	2.5	1.815	8.793	9.357	6.924	5.022	3.762	2.928
0.2,0.8	2.5	1.611	7.597	8.009	5.927	4.327	3.266	2.561
0.2,1	2.5	1.404	6.555	6.900	5.131	3.772	2.867	2.264
0.3,0.3	2.5	1.695	7.513	8.503	6.791	5.104	3.893	3.059
0.3,0.6	2.5	2.320	8.625	8.784	6.408	4.651	3.496	2.731
0.3,0.9	2.5	2.257	7.797	7.650	5.445	3.935	2.964	2.326
0.3,1.2	2.5	1.998	6.709	6.504	4.608	3.341	2.529	1.996
0.3,1.5	2.5	1.750	5.800	5.596	3.966	2.886	2.196	1.742
0.4,0.4	2.5	1.719	6.353	7.003	5.596	4.254	3.284	2.610
0.4,0.8	2.5	2.353	7.273	7.176	5.198	3.797	2.879	2.270
0.4,1.2	2.5	2.287	6.560	6.220	4.377	3.173	2.406	1.902
0.4,1.6	2.5	2.041	5.667	5.291	3.691	2.677	2.036	1.617
0.4,2	2.5	1.772	4.867	4.525	3.156	2.295	1.753	1.397
0.5,0.5	2.5	1.631	5.409	5.890	4.726	3.626	2.827	2.267
0.5,1	2.5	2.241	6.197	6.016	4.350	3.195	2.439	1.936
0.5,1.5	2.5	2.184	5.591	5.203	3.643	2.650	2.019	1.605
0.5,2	2.5	1.943	4.818	4.413	3.059	2.223	1.697	1.353
0.5,2.5	2.5	1.696	4.149	3.778	2.613	1.901	1.455	1.165
0.6,0.6	2.5	1.511	4.682	5.066	4.082	3.155	2.478	2.000
0.6,1.2	2.5	2.076	5.352	5.151	3.727	2.750	2.111	1.685
0.6,1.8	2.5	2.026	4.830	4.449	3.111	2.269	1.736	1.386
0.6,2.4	2.5	1.801	4.156	3.766	2.604	1.896	1.452	1.162
0.6,3	2.5	1.570	3.576	3.220	2.220	1.617	1.241	0.996
0.7,0.7	2.5	1.386	4.107	4.431	3.585	2.788	2.202	1.788
0.7,1.4	2.5	1.910	4.698	4.498	3.258	2.413	1.860	1.491
0.7,2.1	2.5	1.858	4.227	3.873	2.708	1.981	1.521	1.218
0.7,2.8	2.5	1.654	3.640	3.278	2.264	1.651	1.268	1.018
0.7,3.5	2.5	1.445	3.134	2.802	1.928	1.407	1.082	0.870
0.8,0.8	2.5	1.271	3.652	3.933	3.194	2.496	1.981	1.616
0.8,1.6	2.5	1.751	4.172	3.983	2.890	2.147	1.661	1.336
0.8,2.4	2.5	1.705	3.755	3.427	2.397	1.758	1.353	1.087
0.8,3.2	2.5	1.517	3.230	2.897	2.001	1.462	1.125	0.905
0.8,4	2.5	1.324	2.779	2.475	1.702	1.243	0.958	0.772
0.9,0.9	2.5	1.168	3.283	3.532	2.878	2.258	1.800	1.474
0.9,1.8	2.5	1.610	3.749	3.572	2.596	1.934	1.500	1.210
0.9,2.7	2.5	1.567	3.371	3.070	2.149	1.579	1.217	0.980
0.9,3.6	2.5	1.395	2.900	2.594	1.792	1.311	1.011	0.814
0.9,4.5	2.5	1.215	2.492	2.214	1.522	1.113	0.859	0.693
1,1	2.5	1.077	2.978	3.204	2.618	2.061	1.649	1.354
1,2	2.5	1.485	3.400	3.236	2.355	1.759	1.368	1.106
1,3	2.5	1.445	3.056	2.779	1.947	1.432	1.107	0.893
1,4	2.5	1.286	2.629	2.347	1.622	1.188	0.917	0.740
1,5	2.5	1.121	2.259	2.002	1.377	1.008	0.779	0.629

Table C.3: Values of TPA in Tbps/cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1	0.905	10.551	21.314	32.716	37.108	38.305	38.261
0.1,0.2	1	0.677	7.992	16.587	26.674	31.284	33.092	33.670
0.1,0.3	1	0.517	6.238	13.261	22.130	26.658	28.771	29.735
0.1,0.4	1	0.416	5.098	11.017	18.874	23.190	25.416	26.594
0.1,0.5	1	0.348	4.307	9.418	16.448	20.514	22.756	24.048
0.2,0.2	1	1.755	19.140	37.082	54.050	58.972	58.923	57.205
0.2,0.4	1	1.303	14.125	27.681	41.589	46.631	47.668	47.167
0.2,0.6	1	0.993	10.865	21.579	33.217	38.008	39.521	39.676
0.2,0.8	1	0.797	8.791	17.625	27.587	32.017	33.699	34.190
0.2,1	1	0.665	7.377	14.888	23.580	27.648	29.364	30.030
0.3,0.3	1	2.443	23.007	41.184	56.060	59.393	58.466	56.297
0.3,0.6	1	1.812	16.860	30.300	42.042	45.468	45.614	44.669
0.3,0.9	1	1.378	12.896	23.386	33.025	36.285	36.915	36.603
0.3,1.2	1	1.105	10.400	18.983	27.132	30.134	30.956	30.964
0.3,1.5	1	0.922	8.709	15.969	23.018	25.761	26.649	26.826
0.4,0.4	1	2.948	23.763	39.840	51.829	54.161	53.120	51.167
0.4,0.8	1	2.185	17.361	29.089	38.289	40.640	40.485	39.573
0.4,1.2	1	1.661	13.244	22.337	29.802	32.038	32.292	31.904
0.4,1.6	1	1.332	10.663	18.073	24.340	26.394	26.817	26.691
0.4,2	1	1.111	8.917	15.167	20.563	22.435	22.924	22.937
0.5,0.5	1	3.282	23.006	36.789	46.559	48.376	47.473	45.859
0.5,1	1	2.431	16.771	26.722	34.045	35.799	35.590	34.822
0.5,1.5	1	1.848	12.777	20.458	26.347	28.000	28.115	27.766
0.5,2	1	1.481	10.274	16.517	21.437	22.947	23.199	23.056
0.5,2.5	1	1.235	8.586	13.842	18.063	19.434	19.741	19.709
0.6,0.6	1	3.482	21.688	33.536	41.717	43.250	42.523	41.214
0.6,1.2	1	2.577	15.779	24.261	30.273	31.678	31.488	30.862
0.6,1.8	1	1.962	12.028	18.557	23.353	24.653	24.716	24.422
0.6,2.4	1	1.573	9.667	14.963	18.953	20.132	20.303	20.173
0.6,3	1	1.311	8.075	12.529	15.942	17.008	17.223	17.180
0.7,0.7	1	3.561	20.144	30.435	37.475	38.850	38.295	37.241
0.7,1.4	1	2.644	14.678	22.000	27.083	28.269	28.119	27.612
0.7,2.1	1	2.009	11.164	16.783	20.814	21.895	21.945	21.708
0.7,2.8	1	1.611	8.971	13.523	16.864	17.835	17.969	17.862
0.7,3.5	1	1.343	7.490	11.314	14.166	15.039	15.207	15.169
0.8,0.8	1	3.568	18.660	27.720	33.907	35.174	34.757	33.904
0.8,1.6	1	2.649	13.586	19.998	24.402	25.443	25.335	24.925
0.8,2.4	1	2.013	10.329	15.238	18.711	19.643	19.693	19.502
0.8,3.2	1	1.612	8.289	12.260	15.131	15.962	16.078	15.993
0.8,4	1	1.345	6.923	10.258	12.703	13.445	13.585	13.556
0.9,0.9	1	3.522	17.285	25.362	30.890	32.078	31.772	31.076
0.9,1.8	1	2.615	12.578	18.268	22.157	23.094	23.022	22.689
0.9,2.7	1	1.987	9.559	13.908	16.960	17.784	17.837	17.684
0.9,3.6	1	1.593	7.674	11.189	13.704	14.432	14.536	14.470
0.9,4.5	1	1.327	6.404	9.352	11.492	12.139	12.262	12.241
1,1	1	3.442	16.039	23.319	28.324	29.449	29.230	28.660
1,2	1	2.551	11.649	16.755	20.246	21.106	21.068	20.797
1,3	1	1.942	8.861	12.759	15.485	16.228	16.286	16.164
1,4	1	1.556	7.112	10.261	12.501	13.152	13.250	13.199
1,5	1	1.294	5.923	8.562	10.466	11.045	11.158	11.145

[continued] Table C.3: Values of TPA in Tbps/cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1.5	1.290	14.860	29.552	44.102	48.851	49.423	48.518
0.1,0.2	1.5	1.074	12.071	24.196	37.218	42.304	43.658	43.522
0.1,0.3	1.5	0.849	9.627	19.638	31.186	36.324	38.195	38.641
0.1,0.4	1.5	0.690	7.913	16.383	26.670	31.663	33.798	34.609
0.1,0.5	1.5	0.578	6.695	14.019	23.256	28.023	30.274	31.306
0.2,0.2	1.5	2.441	24.970	46.273	64.098	67.840	66.394	63.485
0.2,0.4	1.5	2.019	19.861	36.577	51.352	55.365	55.129	53.520
0.2,0.6	1.5	1.592	15.626	28.993	41.486	45.528	46.040	45.298
0.2,0.8	1.5	1.290	12.723	23.790	34.562	38.444	39.335	39.100
0.2,1	1.5	1.080	10.695	20.123	29.570	33.223	34.296	34.361
0.3,0.3	1.5	3.285	27.860	47.398	61.679	63.919	62.079	59.229
0.3,0.6	1.5	2.715	22.077	37.046	48.262	50.569	49.755	48.080
0.3,0.9	1.5	2.137	17.287	29.101	38.372	40.732	40.571	39.649
0.3,1.2	1.5	1.731	14.034	23.740	31.632	33.916	34.094	33.601
0.3,1.5	1.5	1.448	11.773	19.997	26.859	29.013	29.366	29.124
0.4,0.4	1.5	3.818	27.193	43.568	54.796	56.430	54.884	52.578
0.4,0.8	1.5	3.157	21.530	33.852	42.274	43.781	42.985	41.611
0.4,1.2	1.5	2.485	16.835	26.482	33.323	34.849	34.556	33.768
0.4,1.6	1.5	2.019	13.677	21.572	27.342	28.811	28.779	28.318
0.4,2	1.5	1.688	11.458	18.129	23.119	24.506	24.615	24.347
0.5,0.5	1.5	4.092	25.271	38.944	48.110	49.517	48.343	46.548
0.5,1	1.5	3.385	20.000	30.140	36.763	37.905	37.255	36.176
0.5,1.5	1.5	2.672	15.652	23.546	28.843	29.957	29.679	29.050
0.5,2	1.5	2.164	12.680	19.117	23.556	24.621	24.546	24.169
0.5,2.5	1.5	1.810	10.617	16.045	19.869	20.868	20.901	20.671
0.6,0.6	1.5	4.184	23.106	34.712	42.470	43.775	42.912	41.516
0.6,1.2	1.5	3.463	18.279	26.786	32.229	33.182	32.676	31.828
0.6,1.8	1.5	2.733	14.291	20.882	25.183	26.075	25.849	25.353
0.6,2.4	1.5	2.212	11.566	16.927	20.511	21.352	21.280	20.980
0.6,3	1.5	1.852	9.689	14.204	17.278	18.058	18.068	17.881
0.7,0.7	1.5	4.154	21.036	31.073	37.823	39.071	38.448	37.356
0.7,1.4	1.5	3.447	16.660	23.951	28.575	29.414	29.023	28.350
0.7,2.1	1.5	2.710	12.986	18.610	22.234	22.996	22.823	22.429
0.7,2.8	1.5	2.198	10.519	15.086	18.088	18.790	18.733	18.494
0.7,3.5	1.5	1.838	8.801	12.642	15.210	15.857	15.864	15.714
0.8,0.8	1.5	4.054	19.194	28.021	34.012	35.215	34.772	33.907
0.8,1.6	1.5	3.362	15.182	21.545	25.579	26.347	26.050	25.510
0.8,2.4	1.5	2.647	11.841	16.735	19.869	20.540	20.409	20.092
0.8,3.2	1.5	2.147	9.589	13.556	16.141	16.750	16.709	16.516
0.8,4	1.5	1.795	8.022	11.356	13.561	14.116	14.125	14.004
0.9,0.9	1.5	3.919	17.595	25.470	30.866	32.025	31.716	31.023
0.9,1.8	1.5	3.255	13.926	19.569	23.148	23.856	23.627	23.185
0.9,2.7	1.5	2.567	10.868	15.197	17.956	18.557	18.455	18.194
0.9,3.6	1.5	2.079	8.789	12.294	14.563	15.102	15.073	14.915
0.9,4.5	1.5	1.737	7.350	10.293	12.225	12.712	12.724	12.624
1,1	1.5	3.760	16.192	23.298	28.216	29.336	29.129	28.570
1,2	1.5	3.126	12.817	17.884	21.108	21.772	21.598	21.232
1,3	1.5	2.464	9.996	13.876	16.348	16.899	16.823	16.608
1,4	1.5	1.995	8.081	11.218	13.246	13.733	13.717	13.586
1,5	1.5	1.668	6.758	9.390	11.113	11.551	11.566	11.484

[continued] Table C.3: Values of TPA in Tbps/cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2	1.558	18.002	35.601	52.291	57.059	56.985	55.325
0.1,0.2	2	1.407	15.494	30.464	45.540	50.666	51.406	50.537
0.1,0.3	2	1.159	12.673	25.178	38.626	43.914	45.325	45.168
0.1,0.4	2	0.959	10.533	21.168	33.201	38.430	40.237	40.566
0.1,0.5	2	0.809	8.952	18.171	29.013	34.068	36.089	36.737
0.2,0.2	2	2.878	28.228	50.999	68.826	71.809	69.628	66.140
0.2,0.4	2	2.585	23.889	42.354	57.152	60.287	59.186	56.885
0.2,0.6	2	2.124	19.325	34.279	46.838	50.125	49.876	48.514
0.2,0.8	2	1.754	15.924	28.378	39.258	42.523	42.774	42.011
0.2,1	2	1.478	13.452	24.089	33.668	36.816	37.350	36.965
0.3,0.3	2	3.754	29.718	49.317	63.091	64.945	62.848	59.827
0.3,0.6	2	3.379	25.158	40.632	51.271	52.925	51.614	49.582
0.3,0.9	2	2.776	20.301	32.650	41.399	43.135	42.491	41.218
0.3,1.2	2	2.292	16.693	26.889	34.347	36.092	35.849	35.046
0.3,1.5	2	1.932	14.079	22.740	29.243	30.939	30.929	30.419
0.4,0.4	2	4.219	27.853	43.838	54.711	56.261	54.706	52.410
0.4,0.8	2	3.802	23.594	35.941	43.860	44.977	43.914	42.355
0.4,1.2	2	3.117	18.988	28.734	35.086	36.204	35.622	34.634
0.4,1.6	2	2.579	15.622	23.627	28.969	30.073	29.780	29.136
0.4,2	2	2.174	13.162	19.939	24.564	25.634	25.515	25.087
0.5,0.5	2	4.384	25.192	38.405	47.398	48.873	47.796	46.085
0.5,1	2	3.948	21.317	31.347	37.621	38.537	37.741	36.564
0.5,1.5	2	3.246	17.179	25.035	29.959	30.802	30.342	29.587
0.5,2	2	2.678	14.098	20.521	24.624	25.438	25.191	24.696
0.5,2.5	2	2.256	11.868	17.292	20.825	21.603	21.484	21.150
0.6,0.6	2	4.371	22.641	33.831	41.533	42.971	42.243	40.956
0.6,1.2	2	3.942	19.161	27.543	32.744	33.557	32.962	32.057
0.6,1.8	2	3.239	15.426	21.949	25.967	26.667	26.313	25.729
0.6,2.4	2	2.673	12.654	17.970	21.290	21.945	21.749	21.363
0.6,3	2	2.259	10.673	15.154	17.994	18.606	18.503	18.238
0.7,0.7	2	4.253	20.386	30.070	36.831	38.236	37.759	36.779
0.7,1.4	2	3.843	17.266	24.445	28.900	29.648	29.202	28.492
0.7,2.1	2	3.164	13.908	19.467	22.863	23.472	23.196	22.733
0.7,2.8	2	2.614	11.414	15.933	18.718	19.270	19.113	18.805
0.7,3.5	2	2.202	9.603	13.406	15.783	16.295	16.212	16.000
0.8,0.8	2	4.078	18.441	26.968	33.011	34.381	34.085	33.333
0.8,1.6	2	3.689	15.617	21.887	25.801	26.506	26.171	25.606
0.8,2.4	2	3.038	12.578	17.415	20.367	20.918	20.706	20.334
0.8,3.2	2	2.510	10.319	14.244	16.651	17.139	17.017	16.768
0.8,4	2	2.112	8.673	11.972	14.021	14.469	14.405	14.234
0.9,0.9	2	3.881	16.783	24.400	29.874	31.203	31.039	30.457
0.9,1.8	2	3.513	14.211	19.777	23.275	23.945	23.694	23.237
0.9,2.7	2	2.895	11.447	15.728	18.344	18.851	18.687	18.384
0.9,3.6	2	2.391	9.386	12.853	14.978	15.419	15.325	15.121
0.9,4.5	2	2.012	7.889	10.801	12.604	13.003	12.955	12.815
1,1	2	3.680	15.371	22.254	27.263	28.549	28.480	28.027
1,2	2	3.333	13.015	18.020	21.186	21.825	21.636	21.262
1,3	2	2.746	10.479	14.317	16.672	17.145	17.018	16.767
1,4	2	2.269	8.593	11.698	13.603	14.006	13.934	13.765
1,5	2	1.910	7.222	9.828	11.440	11.802	11.766	11.649

[continued] Table C.3: Values of TPA in Tbps/cm² in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2.5	1.749	20.292	40.003	58.127	62.777	62.147	59.889
0.1,0.2	2.5	1.675	18.279	35.517	52.041	57.008	57.139	55.618
0.1,0.3	2.5	1.433	15.330	29.900	44.703	49.904	50.794	50.060
0.1,0.4	2.5	1.211	12.909	25.373	38.667	43.883	45.271	45.112
0.1,0.5	2.5	1.033	11.044	21.883	33.893	38.994	40.683	40.921
0.2,0.2	2.5	3.157	29.901	53.131	70.709	73.295	70.794	67.073
0.2,0.4	2.5	3.009	26.589	45.981	60.561	63.078	61.436	58.721
0.2,0.6	2.5	2.569	22.094	37.987	50.347	53.029	52.241	50.464
0.2,0.8	2.5	2.166	18.466	31.780	42.502	45.233	45.002	43.863
0.2,1	2.5	1.847	15.714	27.124	36.584	39.272	39.385	38.669
0.3,0.3	2.5	3.998	30.041	49.207	62.642	64.467	62.415	59.450
0.3,0.6	2.5	3.821	26.751	42.281	52.525	53.864	52.336	50.155
0.3,0.9	2.5	3.264	22.210	34.724	43.053	44.409	43.490	42.024
0.3,1.2	2.5	2.748	18.512	28.888	35.964	37.350	36.844	35.856
0.3,1.5	2.5	2.349	15.758	24.598	30.761	32.128	31.876	31.195
0.4,0.4	2.5	4.366	27.332	42.737	53.459	55.166	53.791	51.645
0.4,0.8	2.5	4.174	24.341	36.527	44.205	45.206	44.077	42.479
0.4,1.2	2.5	3.566	20.195	29.887	35.929	36.832	36.108	35.024
0.4,1.6	2.5	3.013	16.864	24.850	29.886	30.767	30.323	29.576
0.4,2	2.5	2.564	14.302	21.072	25.423	26.288	26.030	25.506
0.5,0.5	2.5	4.430	24.288	36.980	45.948	47.642	46.778	45.236
0.5,1	2.5	4.249	21.665	31.527	37.668	38.546	37.736	36.554
0.5,1.5	2.5	3.635	17.981	25.749	30.459	31.170	30.625	29.814
0.5,2	2.5	3.069	14.995	21.359	25.235	25.896	25.549	24.986
0.5,2.5	2.5	2.618	12.734	18.114	21.431	22.063	21.845	21.444
0.6,0.6	2.5	4.330	21.575	32.325	40.072	41.745	41.231	40.112
0.6,1.2	2.5	4.155	19.235	27.473	32.618	33.439	32.861	31.970
0.6,1.8	2.5	3.560	15.968	22.409	26.281	26.897	26.490	25.871
0.6,2.4	2.5	3.004	13.307	18.562	21.715	22.264	21.997	21.565
0.6,3	2.5	2.562	11.293	15.725	18.408	18.918	18.748	18.437
0.7,0.7	2.5	4.143	19.259	28.570	35.414	37.052	36.781	35.962
0.7,1.4	2.5	3.984	17.184	24.248	28.693	29.473	29.057	28.370
0.7,2.1	2.5	3.409	14.244	19.735	23.038	23.598	23.293	22.810
0.7,2.8	2.5	2.880	11.874	16.340	19.007	19.486	19.282	18.941
0.7,3.5	2.5	2.458	10.080	13.840	16.096	16.531	16.398	16.152
0.8,0.8	2.5	3.924	17.328	25.538	31.680	33.271	33.167	32.563
0.8,1.6	2.5	3.775	15.455	21.635	25.563	26.310	26.011	25.472
0.8,2.4	2.5	3.233	12.814	17.597	20.484	21.002	20.770	20.385
0.8,3.2	2.5	2.730	10.675	14.557	16.874	17.305	17.147	16.874
0.8,4	2.5	2.329	9.058	12.321	14.274	14.659	14.555	14.357
0.9,0.9	2.5	3.697	15.709	23.053	28.631	30.167	30.181	29.736
0.9,1.8	2.5	3.561	14.011	19.505	23.030	23.746	23.532	23.102
0.9,2.7	2.5	3.049	11.611	15.849	18.420	18.905	18.728	18.416
0.9,3.6	2.5	2.576	9.673	13.106	15.158	15.554	15.431	15.208
0.9,4.5	2.5	2.196	8.203	11.085	12.810	13.159	13.078	12.915
1,1	2.5	3.477	14.344	20.987	26.101	27.580	27.677	27.350
1,2	2.5	3.351	12.792	17.740	20.942	21.628	21.476	21.129
1,3	2.5	2.870	10.598	14.405	16.725	17.183	17.046	16.790
1,4	2.5	2.424	8.828	11.906	13.751	14.118	14.022	13.836
1,5	2.5	2.067	7.485	10.067	11.614	11.934	11.870	11.734

Table C.4: Values of latency in ns in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1	49.614	16.079	12.129	10.558	10.393	10.579	10.907
0.1,0.2	1	44.850	15.055	11.573	10.236	10.149	10.375	10.725
0.1,0.3	1	44.351	14.948	11.515	10.202	10.123	10.353	10.706
0.1,0.4	1	44.274	14.932	11.506	10.197	10.119	10.350	10.703
0.1,0.5	1	44.274	14.932	11.506	10.197	10.119	10.350	10.703
0.2,0.2	1	13.044	4.818	4.032	4.043	4.405	4.855	5.341
0.2,0.4	1	11.793	4.498	3.828	3.897	4.279	4.739	5.230
0.2,0.6	1	11.652	4.462	3.805	3.881	4.264	4.726	5.217
0.2,0.8	1	11.639	4.458	3.803	3.879	4.263	4.724	5.216
0.2,1	1	11.643	4.459	3.804	3.879	4.264	4.725	5.217
0.3,0.3	1	6.272	2.733	2.533	2.837	3.297	3.796	4.310
0.3,0.6	1	5.662	2.540	2.392	2.722	3.191	3.694	4.212
0.3,0.9	1	5.597	2.520	2.378	2.710	3.180	3.684	4.201
0.3,1.2	1	5.590	2.518	2.376	2.709	3.178	3.682	4.200
0.3,1.5	1	5.589	2.518	2.376	2.709	3.178	3.682	4.200
0.4,0.4	1	3.900	2.003	2.008	2.415	2.909	3.425	3.950
0.4,0.8	1	3.518	1.856	1.890	2.311	2.810	3.329	3.856
0.4,1.2	1	3.478	1.841	1.878	2.301	2.800	3.319	3.846
0.4,1.6	1	3.473	1.839	1.877	2.299	2.799	3.318	3.844
0.4,2	1	3.473	1.839	1.877	2.299	2.799	3.318	3.845
0.5,0.5	1	2.801	1.665	1.765	2.219	2.729	3.253	3.783
0.5,1	1	2.526	1.539	1.658	2.121	2.635	3.160	3.691
0.5,1.5	1	2.496	1.526	1.647	2.111	2.624	3.150	3.681
0.5,2	1	2.494	1.525	1.646	2.110	2.623	3.149	3.680
0.5,2.5	1	2.494	1.525	1.646	2.110	2.624	3.149	3.680
0.6,0.6	1	2.198	1.478	1.630	2.110	2.628	3.157	3.689
0.6,1.2	1	1.983	1.365	1.530	2.017	2.537	3.067	3.600
0.6,1.8	1	1.955	1.351	1.518	2.004	2.526	3.055	3.588
0.6,2.4	1	1.953	1.350	1.516	2.003	2.525	3.054	3.587
0.6,3	1	1.953	1.350	1.516	2.003	2.525	3.054	3.587
0.7,0.7	1	1.841	1.369	1.552	2.047	2.571	3.102	3.636
0.7,1.4	1	1.656	1.260	1.453	1.953	2.478	3.010	3.544
0.7,2.1	1	1.635	1.248	1.442	1.943	2.468	3.000	3.535
0.7,2.8	1	1.633	1.247	1.441	1.941	2.467	2.999	3.533
0.7,3.5	1	1.633	1.247	1.441	1.941	2.467	2.999	3.533
0.8,0.8	1	1.608	1.297	1.500	2.005	2.532	3.065	3.600
0.8,1.6	1	1.445	1.193	1.403	1.912	2.441	2.974	3.510
0.8,2.4	1	1.428	1.182	1.393	1.902	2.431	2.964	3.500
0.8,3.2	1	1.427	1.181	1.393	1.902	2.431	2.964	3.499
0.8,4	1	1.426	1.181	1.392	1.901	2.430	2.963	3.499
0.9,0.9	1	1.449	1.248	1.465	1.977	2.506	3.040	3.576
0.9,1.8	1	1.301	1.147	1.370	1.885	2.415	2.950	3.486
0.9,2.7	1	1.285	1.136	1.359	1.875	2.405	2.940	3.476
0.9,3.6	1	1.283	1.135	1.358	1.874	2.404	2.939	3.475
0.9,4.5	1	1.284	1.135	1.359	1.874	2.405	2.939	3.475
1,1	1	1.335	1.213	1.440	1.957	2.488	3.023	3.559
1,2	1	1.201	1.116	1.348	1.867	2.399	2.934	3.470
1,3	1	1.184	1.104	1.336	1.856	2.388	2.923	3.459
1,4	1	1.182	1.102	1.335	1.854	2.387	2.922	3.458
1,5	1	1.185	1.105	1.337	1.857	2.389	2.924	3.460

[continued] Table C.4: Values of latency in ns in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	1.5	34.798	11.393	8.709	7.771	7.818	8.110	8.500
0.1,0.2	1.5	28.250	9.940	7.893	7.273	7.426	7.771	8.193
0.1,0.3	1.5	26.977	9.657	7.734	7.176	7.349	7.705	8.133
0.1,0.4	1.5	26.672	9.589	7.696	7.153	7.331	7.689	8.119
0.1,0.5	1.5	26.612	9.576	7.689	7.149	7.327	7.686	8.116
0.2,0.2	1.5	9.366	3.675	3.206	3.375	3.791	4.267	4.769
0.2,0.4	1.5	7.595	3.178	2.868	3.116	3.558	4.048	4.557
0.2,0.6	1.5	7.249	3.080	2.801	3.065	3.512	4.005	4.516
0.2,0.8	1.5	7.171	3.058	2.786	3.054	3.502	3.995	4.506
0.2,1	1.5	7.153	3.053	2.783	3.051	3.500	3.993	4.504
0.3,0.3	1.5	4.656	2.246	2.187	2.561	3.045	3.556	4.077
0.3,0.6	1.5	3.769	1.925	1.937	2.346	2.841	3.358	3.884
0.3,0.9	1.5	3.599	1.864	1.889	2.305	2.803	3.321	3.847
0.3,1.2	1.5	3.559	1.849	1.878	2.295	2.794	3.312	3.838
0.3,1.5	1.5	3.550	1.846	1.875	2.293	2.791	3.310	3.836
0.4,0.4	1.5	3.005	1.745	1.829	2.275	2.783	3.306	3.835
0.4,0.8	1.5	2.428	1.486	1.610	2.076	2.590	3.116	3.647
0.4,1.2	1.5	2.316	1.436	1.568	2.038	2.553	3.080	3.611
0.4,1.6	1.5	2.283	1.421	1.555	2.026	2.542	3.069	3.600
0.4,2	1.5	2.277	1.418	1.553	2.024	2.540	3.067	3.598
0.5,0.5	1.5	2.242	1.513	1.664	2.143	2.662	3.190	3.723
0.5,1	1.5	1.808	1.283	1.459	1.951	2.474	3.005	3.538
0.5,1.5	1.5	1.720	1.236	1.418	1.912	2.436	2.967	3.501
0.5,2	1.5	1.700	1.225	1.408	1.903	2.427	2.958	3.492
0.5,2.5	1.5	1.695	1.223	1.406	1.901	2.425	2.956	3.490
0.6,0.6	1.5	1.827	1.387	1.574	2.071	2.596	3.127	3.661
0.6,1.2	1.5	1.472	1.173	1.377	1.884	2.411	2.944	3.479
0.6,1.8	1.5	1.400	1.129	1.338	1.846	2.374	2.907	3.442
0.6,2.4	1.5	1.384	1.120	1.329	1.837	2.365	2.899	3.434
0.6,3	1.5	1.378	1.116	1.326	1.834	2.363	2.896	3.431
0.7,0.7	1.5	1.578	1.311	1.520	2.028	2.557	3.090	3.625
0.7,1.4	1.5	1.267	1.105	1.327	1.842	2.372	2.906	3.442
0.7,2.1	1.5	1.209	1.067	1.291	1.807	2.338	2.872	3.408
0.7,2.8	1.5	1.193	1.056	1.281	1.797	2.328	2.863	3.399
0.7,3.5	1.5	1.190	1.054	1.279	1.795	2.326	2.861	3.397
0.8,0.8	1.5	1.416	1.262	1.485	2.000	2.531	3.065	3.601
0.8,1.6	1.5	1.137	1.063	1.296	1.816	2.348	2.884	3.420
0.8,2.4	1.5	1.083	1.025	1.260	1.781	2.313	2.849	3.385
0.8,3.2	1.5	1.069	1.014	1.250	1.771	2.304	2.839	3.376
0.8,4	1.5	1.066	1.012	1.248	1.769	2.302	2.837	3.374
0.9,0.9	1.5	1.303	1.228	1.460	1.980	2.512	3.048	3.584
0.9,1.8	1.5	1.045	1.032	1.273	1.796	2.330	2.866	3.403
0.9,2.7	1.5	0.994	0.993	1.235	1.760	2.294	2.830	3.367
0.9,3.6	1.5	0.982	0.984	1.227	1.752	2.286	2.822	3.359
0.9,4.5	1.5	0.979	0.982	1.225	1.750	2.284	2.820	3.357
1,1	1.5	1.224	1.204	1.443	1.967	2.500	3.036	3.572
1,2	1.5	0.980	1.011	1.257	1.783	2.318	2.854	3.391
1,3	1.5	0.932	0.973	1.220	1.747	2.282	2.818	3.356
1,4	1.5	0.921	0.964	1.212	1.739	2.274	2.810	3.347
1,5	1.5	0.919	0.962	1.210	1.737	2.272	2.808	3.345

[continued] Table C.4: Values of latency in ns in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2	28.814	9.391	7.206	6.517	6.646	6.979	7.395
0.1,0.2	2	21.545	7.727	6.242	5.903	6.149	6.541	6.991
0.1,0.3	2	19.756	7.317	6.004	5.752	6.026	6.433	6.892
0.1,0.4	2	19.181	7.185	5.928	5.703	5.987	6.398	6.860
0.1,0.5	2	18.991	7.142	5.903	5.687	5.974	6.386	6.849
0.2,0.2	2	7.938	3.243	2.898	3.130	3.565	4.052	4.560
0.2,0.4	2	5.926	2.630	2.460	2.779	3.244	3.746	4.262
0.2,0.6	2	5.425	2.477	2.351	2.692	3.164	3.669	4.188
0.2,0.8	2	5.268	2.430	2.317	2.664	3.139	3.645	4.164
0.2,1	2	5.215	2.413	2.305	2.655	3.130	3.637	4.156
0.3,0.3	2	4.069	2.103	2.100	2.501	2.994	3.510	4.034
0.3,0.6	2	3.022	1.682	1.757	2.198	2.704	3.226	3.754
0.3,0.9	2	2.764	1.578	1.672	2.123	2.632	3.155	3.685
0.3,1.2	2	2.682	1.545	1.645	2.099	2.609	3.133	3.663
0.3,1.5	2	2.653	1.534	1.636	2.091	2.601	3.125	3.655
0.4,0.4	2	2.716	1.705	1.820	2.282	2.794	3.320	3.850
0.4,0.8	2	2.011	1.352	1.512	1.995	2.515	3.044	3.577
0.4,1.2	2	1.842	1.267	1.437	1.926	2.448	2.978	3.511
0.4,1.6	2	1.783	1.237	1.411	1.902	2.424	2.955	3.488
0.4,2	2	1.764	1.228	1.403	1.894	2.417	2.947	3.481
0.5,0.5	2	2.091	1.521	1.692	2.181	2.703	3.233	3.766
0.5,1	2	1.548	1.202	1.401	1.904	2.430	2.963	3.497
0.5,1.5	2	1.413	1.122	1.328	1.835	2.362	2.895	3.430
0.5,2	2	1.371	1.097	1.305	1.813	2.341	2.874	3.409
0.5,2.5	2	1.356	1.089	1.298	1.806	2.334	2.867	3.402
0.6,0.6	2	1.749	1.420	1.621	2.125	2.652	3.184	3.719
0.6,1.2	2	1.292	1.118	1.338	1.852	2.382	2.916	3.452
0.6,1.8	2	1.179	1.043	1.268	1.785	2.316	2.850	3.386
0.6,2.4	2	1.144	1.020	1.246	1.763	2.295	2.829	3.365
0.6,3	2	1.128	1.009	1.237	1.754	2.285	2.820	3.356
0.7,0.7	2	1.543	1.358	1.577	2.090	2.620	3.154	3.690
0.7,1.4	2	1.136	1.066	1.299	1.820	2.352	2.887	3.424
0.7,2.1	2	1.035	0.993	1.230	1.752	2.285	2.821	3.357
0.7,2.8	2	1.002	0.970	1.208	1.731	2.264	2.799	3.336
0.7,3.5	2	0.992	0.962	1.201	1.723	2.257	2.792	3.329
0.8,0.8	2	1.410	1.319	1.550	2.069	2.601	3.136	3.672
0.8,1.6	2	1.037	1.034	1.275	1.800	2.334	2.869	3.406
0.8,2.4	2	0.944	0.962	1.207	1.733	2.267	2.803	3.340
0.8,3.2	2	0.914	0.939	1.185	1.711	2.245	2.782	3.319
0.8,4	2	0.905	0.933	1.178	1.705	2.239	2.775	3.312
0.9,0.9	2	1.319	1.292	1.531	2.054	2.587	3.123	3.660
0.9,1.8	2	0.969	1.011	1.259	1.786	2.321	2.857	3.394
0.9,2.7	2	0.881	0.941	1.191	1.719	2.254	2.791	3.328
0.9,3.6	2	0.854	0.919	1.169	1.698	2.233	2.770	3.307
0.9,4.5	2	0.845	0.912	1.163	1.692	2.227	2.763	3.301
1,1	2	1.254	1.273	1.518	2.044	2.578	3.114	3.651
1,2	2	0.920	0.996	1.247	1.777	2.312	2.849	3.386
1,3	2	0.837	0.926	1.180	1.710	2.246	2.782	3.320
1,4	2	0.810	0.904	1.158	1.689	2.224	2.761	3.299
1,5	2	0.802	0.897	1.151	1.682	2.218	2.755	3.292

[continued] Table C.4: Values of latency in ns in design space.

w,s (μm)	AR	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.1,0.1	2.5	25.670	8.323	6.397	5.838	6.010	6.366	6.794
0.1,0.2	2.5	18.100	6.537	5.334	5.136	5.429	5.845	6.309
0.1,0.3	2.5	15.972	6.035	5.035	4.939	5.266	5.698	6.173
0.1,0.4	2.5	15.181	5.849	4.924	4.866	5.205	5.644	6.122
0.1,0.5	2.5	14.866	5.774	4.880	4.837	5.181	5.622	6.102
0.2,0.2	2.5	7.234	3.058	2.778	3.042	3.488	3.981	4.492
0.2,0.4	2.5	5.086	2.356	2.257	2.611	3.088	3.595	4.115
0.2,0.6	2.5	4.482	2.159	2.110	2.490	2.975	3.487	4.009
0.2,0.8	2.5	4.260	2.086	2.057	2.445	2.934	3.447	3.970
0.2,1	2.5	4.169	2.056	2.035	2.427	2.917	3.431	3.955
0.3,0.3	2.5	3.817	2.082	2.107	2.523	3.021	3.539	4.065
0.3,0.6	2.5	2.670	1.579	1.685	2.142	2.653	3.177	3.707
0.3,0.9	2.5	2.347	1.438	1.566	2.034	2.549	3.076	3.607
0.3,1.2	2.5	2.233	1.388	1.524	1.996	2.512	3.040	3.571
0.3,1.5	2.5	2.179	1.364	1.504	1.978	2.495	3.022	3.554
0.4,0.4	2.5	2.623	1.741	1.873	2.342	2.858	3.384	3.916
0.4,0.8	2.5	1.831	1.310	1.487	1.979	2.502	3.033	3.566
0.4,1.2	2.5	1.608	1.189	1.379	1.877	2.402	2.934	3.468
0.4,1.6	2.5	1.523	1.143	1.337	1.838	2.364	2.896	3.431
0.4,2	2.5	1.492	1.126	1.323	1.824	2.350	2.883	3.417
0.5,0.5	2.5	2.070	1.583	1.764	2.258	2.782	3.313	3.847
0.5,1	2.5	1.438	1.183	1.394	1.902	2.431	2.964	3.499
0.5,1.5	2.5	1.260	1.071	1.289	1.802	2.332	2.866	3.402
0.5,2	2.5	1.195	1.029	1.251	1.765	2.295	2.830	3.365
0.5,2.5	2.5	1.167	1.012	1.235	1.750	2.280	2.814	3.350
0.6,0.6	2.5	1.767	1.495	1.704	2.211	2.739	3.273	3.808
0.6,1.2	2.5	1.226	1.115	1.343	1.861	2.392	2.927	3.463
0.6,1.8	2.5	1.073	1.007	1.241	1.762	2.294	2.829	3.366
0.6,2.4	2.5	1.017	0.968	1.204	1.726	2.258	2.794	3.330
0.6,3	2.5	0.994	0.952	1.188	1.711	2.243	2.779	3.316
0.7,0.7	2.5	1.586	1.444	1.668	2.184	2.715	3.249	3.785
0.7,1.4	2.5	1.097	1.073	1.312	1.835	2.368	2.904	3.441
0.7,2.1	2.5	0.961	0.970	1.213	1.738	2.272	2.808	3.345
0.7,2.8	2.5	0.910	0.931	1.176	1.702	2.236	2.772	3.309
0.7,3.5	2.5	0.888	0.915	1.160	1.686	2.221	2.757	3.294
0.8,0.8	2.5	1.469	1.410	1.645	2.166	2.698	3.234	3.770
0.8,1.6	2.5	1.015	1.046	1.293	1.819	2.354	2.890	3.427
0.8,2.4	2.5	0.887	0.945	1.194	1.722	2.257	2.794	3.331
0.8,3.2	2.5	0.840	0.907	1.157	1.686	2.221	2.758	3.295
0.8,4	2.5	0.821	0.891	1.142	1.671	2.207	2.743	3.281
0.9,0.9	2.5	1.388	1.387	1.629	2.154	2.687	3.223	3.760
0.9,1.8	2.5	0.958	1.028	1.279	1.808	2.343	2.880	3.417
0.9,2.7	2.5	0.838	0.928	1.181	1.712	2.247	2.784	3.322
0.9,3.6	2.5	0.793	0.891	1.145	1.676	2.212	2.749	3.286
0.9,4.5	2.5	0.775	0.876	1.130	1.661	2.197	2.734	3.272
1,1	2.5	1.331	1.371	1.618	2.145	2.680	3.216	3.753
1,2	2.5	0.917	1.015	1.269	1.800	2.336	2.873	3.411
1,3	2.5	0.802	0.916	1.172	1.704	2.240	2.777	3.315
1,4	2.5	0.759	0.879	1.136	1.668	2.205	2.742	3.279
1,5	2.5	0.742	0.864	1.122	1.654	2.190	2.728	3.265

REFERENCES

- [1] S. Naffziger and G. Hammond, "The implementation of the next-generation 64b Itanium Microprocessor," *Tech. Dig. IEEE Int. Symp. Solid-State Circuits (ISSCC)*, pp. 471-473, 2002.
- [2] International Technology Roadmap for Semiconductors (ITRS), 2004 update. [Online document], Available HTTP: <http://public.itrs.net>.
- [3] R. Ho, K. Mai, and M. Horowitz, "Managing wire scaling: a circuit perspective," *Proc. IEEE Int. Interconnect Tech. Conf. (IITC)*, 2003, pp. 177-179.
- [4] International Technology Roadmap for Semiconductors (ITRS), 2000 update. [Online document], Available HTTP: <http://public.itrs.net/Files/2000UpdateFinal/2kUdFinal.htm>.
- [5] J. Xu and W. Wolf, "A wave-pieplined on-chip interconnect structure for networks-on-chips," *Proc. 11th Symp. on High Performance Interconencts (HOTI)*, 2003, pp. 10-14.
- [6] H. Bakoglu and J. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Devices*, vol. ED-32, no.5, pp. 903-909, May 1985.
- [7] H. Bakoglu, "*Circuits, interconnections, and packaging for VLSI*," Addison-Wesley publication, 1990.
- [8] J. Eble, V. De, D. Wills, and J. Meindl, "Minimum repeater count, size and energy dissipation for gigascale integration interconnects," *Proc. IEEE Int. Interconnect Technology Conf. (IITC)*, 1998, pp. 56-58.
- [9] A. Naeemi and J. Meindl, "An optimal partition between on-chip and on-board interconnects," *Proc. IEEE Int. Interconnect Technology Conf. (IITC)*, 2001, pp. 131-133.
- [10] D. Sylvester and K. Keutzer, "Impact of small process geometry on microarchitectures in systems on a chip," *Proc. IEEE*, vol. 89, no. 5, pp. 467-489, Apr. 2001.
- [11] G. Garcea, N. Mejis, and R. Otten, "Simultaneous analytic area and power optimization for repeater insertion," *Proc. IEEE Int. Conf. on Computer Aided Design (ICCAD)*, 2003, pp. 568-573.
- [12] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 2001-2007, Nov. 2002.

- [13] V. Adler and E. Friedman, "Repeater design to reduce delay and power in resistive interconnect," *IEEE Trans. Circuits and Systems – II: Analog and Digital signal Processing*, vol. 45, no. 5, pp. 607-616, May 1998.
- [14] M. Ghoneima and Y. Ismail, "Optimum positioning of interleaved repeaters in bidirectional buses," *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, vol. 24, no. 3, pp. 461-469, March 2005.
- [15] V. Chandra, A. Xu, and H. Schmit, "A low-power approach to system level pipelined interconnect design," *Proc. IEEE/ACM Int. Workshop on System Level Interconnect Prediction (SLIP)*, 2004, pp. 45-52.
- [16] L. Zhang, Y. Hu, and C. Chen, "Wave pipelined on-chip global interconnect," accepted for publication in *Proc. Asia and South Pacific Design Automation Conf. (ASP-DAC)*, 2005.
- [17] L. Zhong and N. Jha, "Interconnect-aware high-level synthesis for low power," *Proc. IEEE Int. Conf. on Computer Aided Design (ICCAD)*, 2002, pp. 110-117.
- [18] H. Zhang, V. George, and J. Rabaey, "Low-swing on-chip signaling techniques: effectiveness and robustness," *IEEE Trans. VLSI Systems*, vol. 8, no. 3, pp. 264-272, June 2000.
- [19] Y. Nakagome et al., "Sub-1V swing internal bus architecture for future low-power ULSI's," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 414-419, Apr. 1993.
- [20] M. Stan and W. Burleson, "Bus-invert coding for low-power I/O," *IEEE Trans. VLSI Systems*, vol. 3, no. 1, pp. 49-58, Mar. 1995.
- [21] P. Wang, G. Pei, and E. Kan, "Pulsed wave interconnect," *IEEE Trans. VLSI Systems*, vol. 12, no. 5, pp. 453-463, May 2004.
- [22] J. Butts and G. Sohi, "A static power model for architects," *Proc. 33rd Ann. Int. Symp. on Micro-architecture*, 2000, pp. 191-201.
- [23] R. Gu and M. Elmasry, "Power dissipation analysis and optimization of deep submicron CMOS digital circuits," *IEEE Journal Solid-State Circuits*, vol. 31, no. 5, pp. 707-713, May 1996.
- [24] X. Chen and L. Peh, "Leakage power modeling and optimization in interconnect networks," *Proc. IEEE Int. Symp. Low-Power Electronic Design (ISLPED)*, 2003, pp. 90-95.

- [25] C. Kang, S. Abbaspour, and M. Pedram, "Buffer sizing for minimum energy-delay product by using an approximating polynomial," *Proc. Great Lakes Symp. on VLSI, (GLSVLSI)*, 2003, pp. 112-115.
- [26] L. Bisdounis and O. Koufopavlou, "Analytical modeling of short-circuit energy dissipation in submicron CMOS structures," *Proc. IEEE Int. Conf. on Electronics, Circuits, and Systems (ICECS)*, 1999, pp. 1667-1670.
- [27] T. Theis, "The future of interconnect technology," *IBM Journal of Research and Development*, vol. 44, no. 3, 2000.
- [28] A. Joshi and J. Davis, "Wave-pipelined multiplexed (WPM) routing for gigascale integration (GSI)," to be published in *IEEE Trans. VLSI systems*, Sept. 2005.
- [29] S. Yang et. al., "A high performance 180nm generation logic technology," *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 1998, pp. 197-200.
- [30] M. Bamal, E. Grossar, M. Stucchi, and K. Maex, "Interconnect width selection for deep submicron designs using the table lookup method," *Proc. IEEE/ACM Int. Workshop on System Level Interconnect Prediction (SLIP)*, 2004, pp. 41-44.
- [31] D. Pamunnwa, L. Zheng, and H. Tenhunen, "Maximizing throughput over parallel wire structures in the deep submicrometer regime," *IEEE Trans. VLSI Systems*, vol. 11, no. 2, pp. 224-243, Apr. 2003.
- [32] A. Kahng and D. Stroobandt, "Wiring layer assignments with consistent stage delays," *Proc. IEEE/ACM Int. Workshop on System Level Interconnect Prediction (SLIP)*, 2000, pp. 115-122.
- [33] P. Zarkesh-Ha and J. Meindl, "An integrated architecture for global interconnects in a gigascale system-on-a-chip," *Proc. IEEE VLSI Symp.*, 2000, pp. 194-195.
- [34] R. Venkatesan, J. Davis, K. Bowman, and J. Meindl, "Optimal n -tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Trans. VLSI Systems*, vol. 9, no. 6, pp. 899-912, Dec. 2001.
- [35] A. Naeemi, "Analysis and optimization for global interconnects for gigascale integration (GSI)," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 2003.
- [36] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, no. 1, Jan. 1993.
- [37] J. Davis and J. Meindl, "Compact distributed RLC interconnect models – I and II – single line transient, time delay, and overshoot expressions," *IEEE Trans. Electron Devices*, vol. 47, no. 11, pp. 2068-2087, Nov. 2000.

- [38] D. Gao, A. Yang, and S. Kang, "Modeling and simulation of interconnection delays and crosstalk in high-speed integrated circuits," *IEEE Trans. on Circuits and Systems*, vol. 37, no. 1, pp.1-9, Jan. 1990.
- [39] R. Dutta and M. Sadowska, "Automatic sizing of power/ground networks in VLSI," *Proc. IEEE Design Automation Conf. (DAC)*, 1989, pp. 783–786.
- [40] L. Cotton, "Maximum rate pipelined systems," *Proc. AFIPS Spring Joint Computer Conf.*, 1969, pp. 581-586.
- [41] D. Wong, G. Micheli, and M. Flynn, "Designing high performance digital circuits using wave pipelining: algorithms and practical experiences," *IEEE Trans. on Computer Aided Design (TCAD)*, vol. 12, no. 1, pp. 25-46, Jan. 1993.
- [42] B. Lim and J. Kang, "A self-timed wave pipelined adder using data align method," *Proc. IEEE Asia-Pacific Conf. on ASICs (AP-ASIC)*, 2000, pp. 77-80.
- [43] T. Feng et al., "Reliability modeling and assurance of clockless wave pipeline," *Proc. IEEE Int. Symp. on Defect and Fault Tolerance in VLSI Systems (DFT)*, 2004, pp. 442-450.
- [44] C. Gray, W. Liu, and R. Cavin, "*Wave-pipelining: theory and CMOS implementation*," Kluwer Academic Publishers, 1994.
- [45] H. Shah et al., "Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects," *Proc. IEEE Int. Conf. on Computer Aided Design (ICCAD)*, 2002, pp. 280-284.
- [46] R. Venkatesan, J. Davis, and J. Meindl, "Compact distributed RLC interconnect models – part IV: unified models for time delay, crosstalk, and repeater insertion," *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 1094-1102, Apr. 2003.
- [47] W. Liu, "*MOSFET models for SPICE simulation, including BSIM3v3 and BSIM4*," John Wiley and Sons publication, 2001, pp. 31-33.
- [48] The MOSIS Service, "HSPICE parameters." [Online document], Available HTTP: <http://www.mosis.org>.
- [49] J. Uyemura, "*CMOS logic circuit design*," Kluwer Academic Publishers, 1999, pp. 201.
- [50] P. Zarkesh-Ha, "Global interconnect modeling for a gigascale system-on-a-chip (GSoC)," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 2001.
- [51] J. Davis, "A hierarchy of interconnect limits and opportunities for gigascale integration (GSI)," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 1999.

- [52] N. Mahmoud, M. Ghoneima, and Y. Ismail, "Physical limitations on the bit-rate of on-chip interconnects," *Proc. Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 13-19, 2005.
- [53] R. Venkatesan, "Multilevel interconnect architectures for gigascale integration (GSI)," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 2003.
- [54] Y. I. Ismail and E. G. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits," *Proc. ACM/IEEE Design Automation Conf. (DAC)*, 1999, pp. 721-724.
- [55] I. Hatirnaz and Y. Leblebici, "Twisted differential on-chip interconnect architecture for inductive/capacitive crosstalk noise cancellation," *Proc. IEEE Int. Symp. Circuits and Systems (ISSCC)*, 2004, pp. V 185 – V 188.
- [56] S. Zhao, K. Roy, and C. Koh, "Decoupling capacitance allocation and its application to power supply noise aware floorplanning," *IEEE Trans. on Computer Aided Design of Integrated Circuits and Systems*, vol. 21, no. 1, pp. 81-92, Jan. 2002.
- [57] L. Smith, "Decoupling capacitor calculations for CMOS circuits," *Proc. of IEEE 3rd Topical Meeting of Electrical Performance of Electronic Packaging*, 1994, pp. 101-105.
- [58] H. Chen and D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design," *Proc. IEEE Design Automation Conference (DAC)*, 1997, pp. 638-643.
- [59] Nalamalpu and W. Burleson, "A practical approach to DSM repeater insertion: satisfying delay constraints while minimizing area and power," *Proc. 14th Annu. IEEE Int. ASIC/SOC Conf.*, 2001, pp. 152-156.
- [60] K. Saraswat, "Performance Analysis and Technology of 3-D ICs," *Invited Talk at Sys. Level Interconnect Prediction (SLIP)*, 2000.
- [61] T. Burd, "Energy efficient system design," Ph.D. Thesis, University of California, Berkeley, 1998.
- [62] W. Dally and A. Chang, "The role of custom design in ASIC chip," *Proc. IEEE/ACM 37th Conf. Design Automation (DAC)*, 2000, pp. 643-647.
- [63] K. Lee et al., "A 51mW 1.6GHz on-chip network for low-power heterogeneous SoC platform," *Tech. Dig. IEEE Int. Symp. Solid-State Circuits (ISSCC)*, 2004, pp. 152-161.

- [64] G. Cardarilli, M. Salmeri, A. Salsano, and O. Simonelli, "Low voltage swing circuits for low dissipation busses," *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 1997, pp. 1868-1871.
- [65] J. Choi, "Modeling of power supply noise in large chips using the finite difference time domain method," Ph. D. Thesis, Georgia Institute of Technology, Atlanta, 2002.
- [66] S. Chun, "Methodologies for modeling simultaneous switching noise in multi-layered packages and boards," Ph. D. Thesis, Georgia Institute of Technology, Atlanta, 2002.
- [67] H. Ando et al., "A 1.3GHz fifth generation SPARC64 microprocessor," *Tech. Dig. IEEE Int. Symp. Solid-State Circuits (ISSCC)*, 2003, pp. 246-255.
- [68] A. Jalabert, S. Murali, L. Benini, and G. Micheli, "X-pipes compiler: a tool for instantiating application specific networks on chip," *Proc. of IEEE Conf. on Design, Automation, and Test in Europe (DATE)*, pp. 884-889, 2004.
- [69] L. Carloni, K. McMillan, and A. Sangiovanni-Vincentelli, "Theory of latency-insensitive design," *IEEE Trans. CAD of ICs and Systems*, vol.20, no.9, pp.1059–1076, Sept. 2001.
- [70] D. Patterson and J. Hennessy, "Computer organization and design: the hardware/software interface," Morgan Kauffman Publishers Inc., pp. 476-488, 1998.
- [71] Q. Xiaohai and C. Hongyi, "Discussion on the low-power CMOS latches and flip-flops," *Proc. IEEE Int. Conf. on Solid State and Integrated Circuit Tech. (ICSICT)*, 1998, pp. 477-480.
- [72] S. Naffziger et al., "The implementation of Itanium 2 microprocessor," *IEEE Journal of Solid-State Ciciuts*, vol. 37, no. 11, pp. 1448-1460, Nov. 2002.
- [73] J. Cong, "An interconnect-centric design flow for nanometer technologies," *Proc. of IEEE*, vol. 89, no.4, pp. 505-528, 2001.
- [74] T. Meincke et al. "Globally asynchronous locally synchronous architecture for large high-performance ASICs," *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*, 1999, pp. II512-II515.
- [75] D. Bertozzi and L. Benini, "Xpipes: a network-on-chip architecture for gigascale system-on-chip," *IEEE Circuits and Systems Magazine*, Second Quarter 2004, pp. 18-31.

- [76] The Berkeley Predictive Technology Model (BPTM). [Online document], Available HTTP: <http://www-device.eecs.berkeley.edu/~ptm>.
- [77] P. Bai et al., "A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and $0.57 \mu\text{m}^2$ SRAM cell," *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2004, pp. 657-660.
- [78] S. Thompson et al., "A 90 nm logic technology featuring 50nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and $1 \mu\text{m}^2$ SRAM cell," *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2002, pp. 61-64.
- [79] S. Thompson et al., "130 nm logic technology featuring 60 nm transistors, low-k dielectrics, and Cu interconnects," *Intel Technology Journal*, vol. 6, no. 2, pp. 1-14.

LIST OF PUBLICATIONS

1. **V. Deodhar** and J. Davis, "Voltage Scaling and Repeater Insertion for High-Throughput Low-Power Interconnects," *IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2003, vol. 5, pp. 349-352.
2. **V. Deodhar** and J. Davis, "Optimization of Throughput Performance for Low-Power VLSI Interconnects," *IEEE Trans. VLSI Systems*, vol. 13, no. 3, pp. 308-318, March 2005.
3. J. Davis, A. Naeemi, and **V. Deodhar**, "Compact Interconnect Modeling and Analysis," Tutorial session, *IEEE Int. Symp. Quality Electronic Design (ISQED)*, 2004.
4. **V. Deodhar** and J. Davis, "Voltage Scaling, Wire Sizing and Repeater Insertion Design Rules for Wave-Pipelined VLSI Global Interconnect Circuits," *IEEE Int. Symp. Quality Electronic Design (ISQED)*, 2005, pp. 592-597.
5. **V. Deodhar** and J. Davis, "Designing for Signal Integrity in Wave-Pipelined SoC Global Interconnects," *IEEE System-on-Chip Conf. (SOCC)*, 2005, pp. 207-210.
6. J. Davis, **V. Deodhar**, and A. Joshi, "Design based approaches vs. process/material solutions for interconnects," invited paper, *Advanced Metallization Conf. (AMC)*, 2005.
7. A. Joshi, **V. Deodhar**, and J. Davis, "Low Power Multilevel Interconnect Networks Using Wave-Pipelined Multiplexed (WPM) Routing," to appear in *IEEE VLSI Design Conf.*, 2006.

VITA

Vinita Vasant Deodhar was born in Mumbai, India in October 1979. In 1995, she was ranked First among 300,000 students in Mumbai at the Secondary School Certificate (SSC) Examination. In 2001, she received a Bachelor of Engineering in Electronics from Veermata Jijabai Technological Institute (VJTI), University of Mumbai, with highest honors. At VJTI, She was ranked First in the Electronics Department for all four years of her undergraduate studies. In August 2001, she had the privilege to join Dr. Jeffrey Davis's research group at Georgia Institute of Technology. Over the past four years, she has co-authored six publications in international conferences and refereed journals. Her research interests include modeling and circuit-level design of on-chip interconnect networks and throughput-centric optimization of wave-pipelined on-chip global interconnects to achieve high performance along with low power and area.